B.D. Mishler. 2005. The logic of the data matrix in
phylogenetic analysis. In V.A. Albert (ed.), Parsimony,
Phylogeny, and Genomics, pp. 57-70.   Oxford University Press.

II

Parsimony, character analysis, and optimization of sequence characters

CHAPTER 4

The logic of the data matrix in phylogenetic analysis

Brent D. Mishler

4.1 Introduction

The process of phylogenetic analysis inherently consists of two phases. First a data matrix is assembled, then a phylogenetic tree is inferred from that matrix. There is obviously some feedback between these two phases, yet they remain logically distinct parts of the overall process. One could easily argue that the first phase of phylogenetic analysis is the most important; the tree is basically just a re-representation of the data matrix with no value added. This is especially true from a parsimony viewpoint, the point of which is to maintain an isomorphism between a data matrix and a cladogram. We should be very suspicious of any attempt to add something beyond the data in translating a matrix into a tree!

Paradoxically, despite the logical preeminence of data matrix construction in phylogenetic analysis, by far the greatest effort in phylogenetic theory has been directed at the second phase of analysis, the question of how to turn a data matrix into a tree. Extensive series of publications have been elaborated to attempt to justify such tree building approaches as neighbor-joining, maximum likelihood, and Bayesian inference, while ignoring entirely the nature of the data matrix that must underlie any analysis. The reasons for this asymmetry in research on phylogenetic theory are not entirely clear, but it probably has to do with the fact that the problem of tree building may appear simpler, more clear-cut. Perhaps it is just a matter of research fashions. For whatever reason, relatively little attention has been paid to the assembly of the data matrix, and it is high time to examine this all-important part of systematic research. At stake are each of the logical elements of the data matrix: the rows (what are the terminals?), the columns (what are the characters?), and the individual entries (what are the character states?).

The tree of life is inherently fractal-like in its complexity, which complicates the search for answers to these questions. Look closely at one *lineage* of a phylogeny (defined as a diachronic connection between an ancestor and a descendent) and it dissolves into many smaller lineages, and so on, down to a very fine scale. Thus the nature of both the *terminal units* (TUs; the twigs of the tree in any particular analysis) and the characters (hypotheses of homology, markers that serve as evidence for the past existence of a lineage) change as one goes up and down this 'fractal' scale. Furthermore, there is a tight interrelationship between TUs and character states, since they are reciprocally recognized during the character analysis process.

This chapter will deal with logical issues involving the elements of the data matrix in light of the nested and interrelated nature of TUs and characters. I will argue at the end that if care is taken to construct an appropriate data matrix to address a particular question of relationships at a given level, then simple parsimony analysis is all that is needed to transform the matrix into a tree. Debates over more-complicated models for tree building can then be seen for what they are: attempts to compensate for marginal data.

4.2 What exactly is a terminal branch on a tree (that is, a row in the data matrix)?

People who publish phylogenetic analyses are usually cavalier about what their terminal branches represent. One often sees species or other taxon names, or even geographic designations of populations, attached to terminal branches of published trees without explanation. Larger-scale units might indeed be a well-justified TU, but they need to be justified, not assumed *a priori*. Taxa or populations are never the fundamental things from which phylogenies are actually built. Not even individuals are the TUs (contra Vrana and Wheeler 1992). As was carefully elaborated by Hennig (1966), the fundamental terminal entity in phylogenetics is the *semaphoront*, an instantaneous time slice of an individual organism at some point in its ontogeny. A tube of extracted DNA and its associated museum voucher specimen—a semaphoront—should be considered the ultimate TU.

This realization helps conceptually, but doesn't solve all of the empirical problems that arise in assembling a matrix. In practice, TUs (i.e. rows in a data matrix) are usually not semaphoronts. Especially in larger-scale studies, TUs are usually a complicated assemblage of semaphoronts, and sometimes even include data removed from any connection with its original semaphoront. Many specimens often need to be examined for relevant character information (not all of which can be gathered from all semaphoronts because of their sex, life stage, or state of preservation). Information from the literature or a database such as GenBank is often included in the matrix, based on a taxon identification alone without reference to a voucher specimen. This process of assembly of such composite TUs needs careful examination.

Similar sorts of terminals have been called operational taxonomic units (OTUs) in the past, but I think a refined concept of TUs, as referred to above, is necessary, one designed specifically for phylogenetics. The original concept of OTU was defined by pheneticists as a minimal cluster in a Euclidian distance sense. Cladists need instead to refer to specific, potentially homologous and discrete-state characters in a Manhattan distance sense. An additional flaw of the original concept of OTU is that, by using the word 'taxonomic,' it implies that one can do taxonomy before an analysis is completed. This view, by confounding the logical precedence of analysis before classification, has led to major mistakes in systematics research, both phenetic and cladistic, most acutely in the development of phylogenetic species concepts (see the debates framed in Wheeler and Meier 2000).

So how can we define a TU that is suitable for use in phylogenetics? Epistemologically speaking, *a TU is a set of semaphoronts that are homogeneous for the informative character states currently known* (as explained in detail below). A TU is essentially a pile of semaphoronts that cannot currently be subdivided by character data, and thus it is a pragmatic unit, always subject to change as knowledge of characters progresses. Ontologically speaking, *a TU is taken to represent a time slice of one of the terminal lineages whose relationships are being studied in a particular analysis*.

Why do I say ''in a particular analysis?'' Because the definition of TUs, even for the same group of organisms, may change in analyses at different scales. There unfortunately isn't one fundamental TU suitable for any and all analyses; for several different reasons. Epistemologically speaking, since TUs are dependent on character-state divisions in the characters being employed, they are discovered and defined in the course of character analysis (as discussed in detail below), and of course different characters are useful at different scales of analysis. There is thus a reciprocal relationship between character states and TUs as they are being discovered during character analysis at different levels. Ontologically speaking, larger-scale lineages are usually composed of smaller lineages nested inside them, and the choice of which lineage to represent in a particular analysis depends on the questions begin asked. Furthermore, the lineages at these different levels potentially have different histories; in other words the smaller lineages are not always proper subsets of the larger ones. This is sometimes called the gene tree/species tree distinction (Maddison and Maddison 1992), but that distinction is far too simplified; there are many nested levels of potentially incongruent lineages, not just two (more on this topic later).

Even if one wanted to try to avoid these problems by using only semaphoronts in a data matrix, one would still need to pay attention to the same issues of scale. One would still need to decide conceptually which lineages are being represented by what semaphoronts. It is nearly impossible in practice to use single semaphoronts as terminals rather than compositely coded TUs that have data taken from a number of semaphoronts. For one thing, not all semaphoronts bear all the characters; there may be juvenile specializations or sexual dimorphism present in a lineage. Some specimens will be missing reproductive organs or other key features. Different genes will often be sequenced from different individuals. Furthermore, data are often taken from the literature (e.g. a previously published ultrastructural analysis) or from a database (e.g. another laboratory's gene sequence), in cases where no reference can be made to an original semaphoront (e.g. if no voucher specimen was deposited in a museum). Thus, data are virtually always compiled from studies of different individual organisms considered to represent the same terminal lineage. TUs are nearly always composites in practice; their composition varying depending on the scale of analysis.

This topic obviously touches on the species debate, on which I have some opinions (Mishler and Brandon 1987; Mishler 1999; Mishler and Theriot 2000a, b, c), but which I am attempting to steer clear of in this essay to maintain focus. I am speaking here to how data matrices are made: classification (including naming species) is something that happens *much* later in the process. So, while this is not the place to debate species concepts, I do need to point out that the fractal scaling of nested lineages includes those well below the traditional species level. Thus, species are not somehow different from lineages at any other level; they are not 'privileged' TUs—they simply need to be justified like any other.

In summary, there is never a given, *a priori* set of TUs to begin a phylogenetic analysis with. Certainly, named taxa (including species) should not be taken as TUs without question. TUs need to be constructed during each analysis, and re-checked each time a group is re-studied. They need to be carefully justified and re-justified using character evidence. This causes problems with easy comparison between analyses based on different data, but is an unavoidable fact of life in systematics and needs to be taken into account in such areas as database design (more below).

4.3 What exactly is a character (that is, a column in the data matrix)?

The fundamental activity in phylogenetic systematics is *character analysis* (Neff 1986) in which characters and states are hypothesized, tested, and refined in a reciprocal manner, in concert with the assembly of TUs, as part of the development of a data matrix. In addition to the logical primacy of data matrix construction, there is a temporal primacy as well. It is an established fact that a systematist spends 95% of his/her time gathering and analyzing character data and less than 5% time turning the assembled data matrix into a tree. Character analysis must be the all-important part of the phylogenetic reconstruction process if there is going to be a hope of discovering the history of a group. Fortunately, there have been some clear treatments of the elements of character analysis (Wiley 1981, Farris 1983, Neff 1986), but these were published some time ago and seem to be unknown to many recent workers. Younger systematists would do well to put more energy into investigations of the principles of character analysis and building better matrices, than into ever more complex model building for tree reconstruction, keeping firmly in mind the principle of 'garbage in, garbage out.' No model of the evolutionary process can be brought to bear successfully if the data matrix does not represent cogently argued character and character-state statements.

Before using a tool (characters in this case) it is wise to think carefully about what one is trying to do with the tool. What we are trying to do in phylogenetic analysis is to infer the existence of some past lineage by finding characters that changed state in that lineage and can thus serve as a potential marker for reconstructing that branch in the future (the Hennig Principle). The goal of

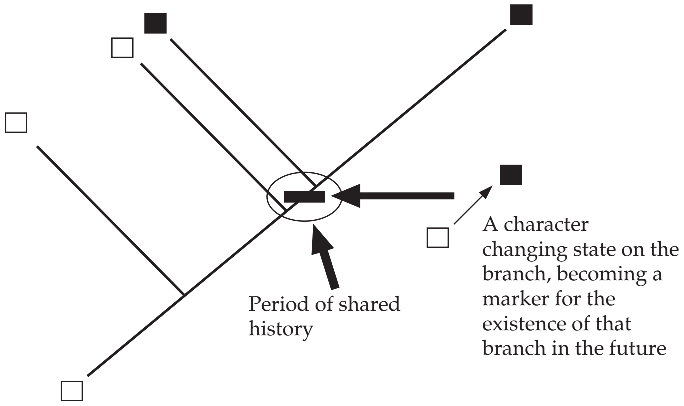
Period of shared history
A character changing state on the branch, becoming a marker for the existence of that branch in the future

Figure 4.1 Illustration of the concept of a phylogenetic marker.

character analysis is find as many potential markers as possible that can serve as evidence for the past existence of lineages shared by one or more of the TUs (see Fig. 4.1). These markers are the only tools a phylogeneticist has to reconstruct the branching history of life, but of course the kind of markers that are useful for branches at one level of depth in time won't necessarily be so for another level. Thus markers need to be searched for carefully in light of the particular branching events one is trying to reconstruct. Since semaphoronts are chosen to build TUs that are representative of the branching events under study, then we need to find 'good' characters that differentiate the chosen semaphoronts.

Much has been written about what constitutes a 'good' character. Ontologically speaking, potentially informative markers need to support a hypothesis of homology across the group under study; thus they need to be comparable in a convincing way across the study organisms. They need to be independent, so they can be taken as separate pieces of evidence for the existence of past lineages in the face of confounding effects such as convergence. They need to have discrete states so they can be inferred to contain a record of evolutionary events marking at least one specific past lineage. The epistemological rules of character analysis can thus be summarized as follows. Potential characters need to be evaluated by evidence for: (1) homology and heritability of a character across the TUs being studied, (2) independent evolution of different characters, and (3) presence in each character of a system of at least two discrete states. I elaborate somewhat on each of these criteria in turn below:

(1) Homology is certainly one of the most important concepts in systematics, and therefore also one of the most controversial. Following on from the work of Hennig and later phylogenetic systematists, when we say that two semaphoronts share the same characteristic, we mean they share a profound historical continuity of information (Roth 1984, 1988). They are postulated to have shared a common ancestor that had that characteristic. Thus an important contribution of cladistics has been the explicit formulation of a phylogenetic criterion for homology: *a hypothesis of taxic homology (i.e. a potential synapomorphy) by necessity is also a hypothesis for the existence of a monophyletic group* (Patterson 1982; Stevens 1984). Each postulated homology (i.e. a column in the data matrix) is essentially a miniature phylogenetic hypothesis all by itself (especially as viewed in the context of its assigned states), and can be tested against other postulated homologies. Therefore, congruence among all postulated homologies provides a test of any single character in question; some characters initially thought to be homologous are later inferred not to be because they are in conflict with the majority of characters. The initial hypotheses of homology are based on detailed similarity in structure and development (see the discussion in Wiley 1981); these go into the matrix for eventual testing by congruence.

(2) For character changes to count as independent pieces of evidence in the congruence test (Patterson 1982), it is necessary that they not be

genetically, developmentally, or functionally correlated with other characters. There are many biological processes acting to distort the phylogenetic signal present in characters (e.g. reversal to primitive states caused by heterochrony, convergent evolution across different characters caused by natural selection, parallel changes to the same state within one character caused by functional constraints, etc.), along with random effects such as long branch attraction (caused by the accumulation of homoplastic matches on long, non-sister branches making them appear to be sister groups). The only weapon the phylogenetic systematist has against this inevitable distortion is many independent sources of information that are, as best as can possibly be determined, not impacted by the *same* biasing factors.

Note that there is another meaning of 'correlation', phylogenetic congruence, that does not disqualify characters from counting as independent! Congruence is what gives us supporting evidence for the existence of a monophyletic group. Thus the rules of character analysis need to be carefully drawn to encompass all the valid potential markers possible while rejecting those that are not suitable.

(3) Why is it necessary for a useful character to have at least two distinct states? Again, we need to think back to what we are trying to do: discrete states are needed because we are trying to reconstruct a discrete thing, an evolutionary event in which a prior state changed to some new posterior state, thus marking the existence of a shared ancestral lineage. The literature on the practice of how to define character states has had a checkered past. In most cases, people have simply made character state distinctions without any justification at all, and many methods proposed for 'gap coding' are flawed in various ways (Stevens 1991). The empirical details are beyond the scope of this chapter; see Mishler and De Luna (1991) for a discussion of this issue and a recommended approach using ANOVA and multiple range tests to seek statistically homogeneous groups of TUs representing character states.

To summarize, a 'good' character for phylogenetic analysis shows greater variation among TUs than within them. This variation must be heritable and independent of other characters. This view of taxonomic characters also requires that each be a system of at least two discrete transformational homologs, or *character states*. Note that just as with TUs, there is never a given, *a priori* set of characters to begin a phylogenetic analysis with. Characters need to be discovered and evaluated during each analysis, and re-checked each time a group is studied.

4.4 What is the relationship between TUs and character states (that is, the individual entries in the data matrix)?

Neither the concept of TU nor the concept of character can be fully understood alone, without reference to each other and to the 'fractal' nature of the tree of life (as discussed earlier). The nature of both TUs and characters change as you go up and down this fractal scale.

As discussed earlier, the rows in a data matrix are virtually never based on data taken from a single individual, given that different labs are producing the data over time, and that different data-gathering techniques (ranging from DNA extraction through preparation for anatomical study) often require destructive sampling; thus data are often compiled from study of different organisms considered to represent the same TU. Thus TUs are nearly always composites in practice, their composition varying depending on the scale of analysis.

Likewise, what counts as a useful character changes depending on the scale of analysis. They have been selected based on their apparent utility for the task at hand, homologized (e.g. aligned) for the organisms under study, and pre-screened for their fit to the criteria of a good taxonomic character. Thus, the columns in a data matrix are already highly refined hypotheses of phylogenetic homology, defined with respect to the scale of the current study.

To make things more complicated, there is clearly a reciprocal relationship between TUs and character states. As detailed earlier, a TU can best be defined as a set of individual samples (semaphoronts in Hennig's terminology) that are homogeneous for character states currently known, while a character can best be defined as a potential marker for shared history of some subset of the known TUs. This means that TUs and characters emerge during a process of ''reciprocal illumination'' (Hennig 1966). To a large extent their definitions and discovery are interlinked. How do we proceed empirically in a way that avoids circularity? Before answering this question we need to consider the scaling problem in more detail.

4.5 Deep vs. shallow phylogenetics

The reconstruction of 'deep' relationships is fundamentally different than reconstruction of 'shallow' relationships (Mishler 2000). This is because the problems faced at these different temporal scales are quite distinct. In shallow reconstruction problems, the branching events at issue happened a relatively short time ago and the set of lineages resulting from these branching events is relatively complete (extinction has not had time to be a major effect). In these situations the relative lengths of internal and external branches are similar, giving less opportunity for long-branch attraction. However, the investigator working at this level has to deal with the potentially confounding effects of reticulation and lineage sorting. Character-state distinctions may be quite subtle, at least at the morphological level. At the nucleotide level it is necessary to look very carefully to find genes evolving rapidly enough; however, such genes may be relatively selectively neutral, and thus less subject to adaptive constraints which can lead to non-independence.

In deep reconstruction problems, the branching events at issue happened a relatively long time ago and the set of lineages resulting from these branching events is relatively incomplete (extinction has had a major effect). In these situations, the relative lengths of internal and external branches are often quite different; thus there is more opportunity for long branch attraction, even though there is little to no problem with reticulation and lineage sorting since most of the remaining branches are so old and widely separated in time. Due to all the time available on many branches, many potential morphological characters should be available, yet they may have changed so much as to make homology assessments difficult; the same is true at the nucleotide level, where multiple substitutions in the same region may make alignment difficult. Thus very slowly evolving genes may be sought, but that very conservatism is caused by strong selective constraints which increases the danger of convergence leading to character dependence. Another approach is to increase sampling density—if TUs can be added that more evenly sample the true tree, thus reducing the asymmetry between internal and external branches, then faster-evolving genes may have better performance (Källersjö *et al.* 1998, 1999).

These considerations suggest that the problems being faced, and their best-justified solutions, will change as you go up and down this fractal scale. The nature of TUs and usable characters are going to change, and we need to have a way to scale phylogenetic results from one level to the next if we are going to have a hope of building a complete tree of life. There is effectively an infinite number of semaphoronts out there; there will never be a 'complete' data matrix for all of them for the practical reason that there are too many. But more importantly, it isn't at all clear that a single global analysis of all semaphoronts living on Earth would be desirable, even if we could do it. There is the fact discussed earlier that a given semaphoront doesn't bear all the relevant data, and thus composite TUs would need to be constructed in practice. There is also the fact that character homologies can be drawn much more easily when comparing only closely related TUs. Very few characters can be coded reliably across the whole tree of life. So we need to examine the scaling issue closely to see how we might combine or concatenate data matrices and phylogenetic results from more-shallow analyses into deeper and deeper ones until eventually a global tree of life can be produced.

4.6 How should we connect up analyses and data matrices that are 'nested' inside each other at various different time scales?

How will we ultimately connect up deep and shallow analyses, each with their own distinctively useful data and problems? Some hold out hope for eventual global analyses, once enough universally comparable data have been gained and computer programs get much more efficient, to deal with all extant organisms at once. Others would go to the opposite extreme, and use a *supertree* approach, where shallow analyses are grafted on to the tips of deeper analyses. An intermediate approach, called *compartmentalization* (Mishler 1994, 2000), uses shallow topologies (that are based on analyses of the characters useful locally) to constrain global deep analyses (that are based on analyses of characters useful globally). All of these issues surrounding how to use phylogenetic markers at their appropriate level to reconstruct the extremely deep tree of life are likely to be among the major concerns of phylogenetics in coming years, as reconstruction of the whole tree of life from twigs to trunk is attempted.

The different approaches to concatenating analyses at different scales can be best viewed as a spectrum (see Fig. 4.2). At the left-hand end of this spectrum, the approach is to include all possible TUs and potential characters in one matrix. Generally this is not actually done, because the sheer amount of data (millions of possible TUs) makes thorough phylogenetic analysis computationally impossible. The most-common approach in practice in global analyses is to select a few representatives of a large, clearly monophyletic group (the *exemplar method*). Care is sometimes taken to select representatives that are 'basal' TUs within the group to be represented (i.e. cladistically basal relative to the imaginary root defined by outgroups); however, this still does not avoid two important problems: (1) within-group variation is not fully represented in the analysis, and (2) an increase both in terminal branch lengths and in asymmetry between lengths of different branches is introduced. These problems can lead to erroneous long branch attractions in global analyses.

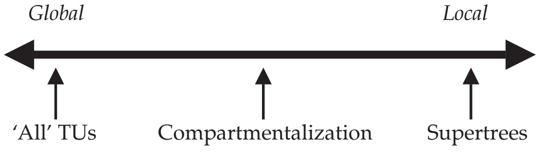

Figure 4.2 How to concatenate different analyses to build the tree of life? Shown is a spectrum of approaches ranging from global to local. See text for explanation.

At the right-hand end of the spectrum, local analyses are simply grafted together into supertrees at the place where shared taxa occur, without reference back to the original data. There are many ways to do this in detail (as reviewed by Sanderson *et al.* 1998), but the important thing is that the analyses on real character data are only done locally, and the concatenation is based on a combination of local topologies rather than an integration of local data sets into a global data set.

Both of these approaches may be problematic, one too global, the other too local. Thus the appeal of a promising intermediate approach called *compartmentalization* (by analogy to a water-tight compartment on a ship—homoplasy is not allowed in or out). This approach represents diverse yet clearly monophyletic clades by their inferred ancestral states in larger-scale cladistic analyses (Mishler 1994, 2000). A well-supported local topology is sought first, then an inferred ''archetype'' or *Hypothetical Ancestor* (HA) for the group is inserted into a more inclusive analysis. In more detail, the procedure is to: (1) perform global analyses, determine the best supported clades (these become the compartments); (2) perform local analyses within compartments, including more taxa and characters (more characters can be used within compartments due to improved homology assessments among closely related organisms—see below); (3) return to a global analyses, in one of two ways, either (a) with compartments represented by single HAs (the archetypes), or (b) with compartments constrained to the topology found in local analyses (for smaller data sets, this approach appears better because it allows flexible character state assignments to the base of the compartment based on optimizations to the local topology).

The compartmentalization approach differs from the exemplar approach in that the representative character-states coded for the archetype are based on all the TUs in the compartment, thus the reconstructed HA is likely to be quite different from any particular TU. As an estimate of the states of the most recent common ancestor of all the local

TUs, the HA is likely to have a much shorter terminal branch with respect to the global analysis, which in turn can have the beneficial global effect of reducing long branch attraction. In addition to these advantages of compartmentalization at the global level, the local analyses will be better because one can: (1) include all local TUs for which data are available; (2) incorporate more (and better justified) characters, by adding in those characters for which homology could not be determined globally (e.g. genes that can only be aligned locally); (3) avoid spurious homoplasy that can change the local topology due to long-branch attractions with distant outgroups. The effects of compartmentalization are thus to cut large data sets down to manageable size, suppress the impact of spurious homoplasy, and allow the use of more information in analyses. This approach is self-reinforcing; as better understanding of phylogeny is gained, the support for compartments will be improved, leading in turn to refined understanding of appropriate characters and TUs both within the compartments *and* between the compartments.

4.7 Structural vs. DNA sequence characters

The choice of data for use at different scales of analysis is the crux of the matter. One important issue to consider is how intrinsically useful are different categories of characters at these different scales? It is clear that, as general categories, structural data (i.e. anatomical, morphological, or genomic) and DNA sequence data have different and complementary strengths and weaknesses. DNA sequence characters are much more numerous than structural characters, thus increasing the chance that sufficient markers can be found for all branches of a tree. They are especially useful in organisms with simple morphology, such as fungi and bacteria, that may lack a sufficient number of structural characters. Objectively defining character states in structural comparisons can be difficult, particularly in shallow reconstructions, while the states are usually clear-cut in DNA sequence data. It has been argued that it is useful that DNA sequence data are independent from morphological characters that are perhaps subject to adaptive convergence (although convergence of course cannot be ruled out in DNA sequences, particularly at deeper levels). Sequences of highly conserved genes can be homologized across very broad groups that share little morphologically, although these same highly conserved regions are probably highly subject to adaptive convergence. Finally, models of evolutionary change are easier to postulate for DNA sequence evolution, a perceived advantage for those who like to use maximum likelihood methods.

On the other hand, especially in deeper comparisons, structural characters (i.e. traditional morphological characters but also modern genomic characters such as rearrangements and intron insertions; see next section) often have much greater complexity, and may exhibit ontogeny, allowing a temporal axis of comparison not available with DNA sequence data. Structural characters often change in an episodic pattern, which is necessary for evidence of deep, short branches to remain detectable. Clock-like markers are the worst kind of data for those sorts of branches; the markers keep changing and thus erasing history. It is much better for discovering those deep, short branches to have a clock like those found frozen in place on the sunken ship *Titanic* (still showing the time the ship went down); a clock that stopped ticking when some major change occurred. Furthermore, the number of possible character states is usually much higher in morphological character systems (and in genomic rearrangements) than in DNA sequence data, which serves to make long branch attraction less of a problem (see Mishler 1994 for discussion). Morphological data are more easily gathered from large numbers of specimens, and from fossils, making it much easier to robustly sample the true phylogeny. For all these reasons, morphological data have remained among the characters of choice at deeper phylogenetic levels, and have been joined recently by an exciting new class of structural characters derived from genomic comparisons. The latter promise to be very useful in the future, particularly for those deep, relatively short internal branches that have proven resistant to phylogenetic reconstruction with DNA sequence data.

4.8 Genomic characters

This is the era of whole-genome sequencing; molecular data are becoming available at a rate unanticipated even a few years ago. Sequencing projects in a number of countries have produced a growing number of fully sequenced genomes, providing computational biologists with tremendous opportunities. However, comparative genomics has so far largely been restricted to pairwise comparisons of genomes; for instance, to identify syntenic regions, orthologous genes, and common regulatory elements between human and mouse. The importance of taking a phylogenetic approach to systematically relating larger sets of genomes has only recently been realized.

A recent synthesis of phylogenetic systematics and molecular biology/genomics—two fields once estranged—is beginning to form a new field that could be called phylogenomics (Eisen *et al.* 1998). Something can be learned about the function of genes by examining them in one organism. However, a much richer array of tools is available using a phylogenetic approach. Close sister-group comparisons between lineages differing in a critical phenotype (e.g. desiccation or freeze tolerance) can allow a quick narrowing of the search for genetic causes. Dissecting a complicated, evolutionarily advanced genotype/phenotype complex (e.g. development of the angiosperm flower) by tracing the components back through simpler ancestral reconstructions can lead to quicker understanding. Hence, phylogenomics allows one to go beyond the use of pairwise sequence similarities and use phylogenic comparative methods to confirm and/or to establish gene function and interactions.

Cross-genome phylogenetic approaches have the potential to provide insights into many open functional questions. A short list includes understanding the processes underlying genomic evolution, identifying key regulatory regions, understanding the complex relationship between phenotype and genomic changes, and understanding the evolution of complex physiological pathways in related organisms. Using such a comparative approach will aid in elucidating how these genes interact to perform specific biological processes. For example, Stuart *et al.* (2003) used microarray data from four completely sequenced genomes (yeast, nematode, insect, and human) to show coexpression relationships that have been conserved across a wide spectrum of animal evolution.

Most importantly for the systematist, the new comparative genomic data should also greatly increase the accuracy of reconstructions of the tree of life. Even though nucleotide sequence comparisons have become the workhorse of phylogenetic analysis at all levels, there are clearly phylogenetic problems for which nucleotide sequence data are poorly suited, because of their simple nature (having only four character states) and tendency to evolve in a regular, more-or-less clock-like fashion. In particular, as stated earlier, deep branching questions (with relatively short internodes of interest mixed with long terminal branches) are notoriously difficult to resolve with DNA sequence data. It is fortunate, therefore, that fundamentally new kinds of structural genomic characters such as inversions, translocations, losses, duplications, and insertion/deletion of introns will be increasingly available in the future.

These characters need to be evaluated using much the same principles of character analysis (discussed earlier) that were originally developed for morphological characters. They must be looked at carefully to establish likely homology (e.g. examining the ends of breakpoints across genomes to see whether a single rearrangement event is likely to have occurred), independence, and discreteness of character states. Given the close link between characters and TUs discussed above, it is also necessary to consider carefully the appropriate TUs for comparative genomic analysis, especially since different parts of one organism's genome may or may not have exactly the same history. Thus close collaboration between systematists and molecular biologists will be required to code these genomic characters properly, and to assemble them into matrices with other data types. Challenges resulting from combining different data sources, in light of the possibility of different histories for different parts of the same genome, are discussed in the next two sections.

4.9 Dealing with heterogeneous data types

There is every reason to search carefully for good potential markers in all kinds of data, particularly for the deep branching questions discussed earlier. Deep phylogenetic reconstructions are by their nature difficult, and all characters should be sought and used if they meet the criteria of good potential markers (Mishler 2000). However, it remains controversial how data from different sources are to be evaluated and compared (Swofford 1991). Some have argued that data sets derived from fundamentally different sources should be analyzed separately, and only common results taken as well supported (i.e. consensus tree approaches), or at least that only data sets that appear to be similar in the trees they favor should be combined (Huelsenbeck *et al.* 1996). Others have argued that all putative homologies should be combined into one matrix (i.e. 'total evidence'; Kluge 1989; Barrett *et al.* 1991; Donoghue and Sanderson 1992; Mishler 1994). Theoretical arguments at present favor the latter approach: if characters have been independently judged to be good candidates for phylogenetic markers, then they are equivalent and should be analyzed together.

There is one major exception to the preference for a 'total evidence' position: data should not be combined into a single matrix if there is evidence that some characters had a different branching history than the rest (Mishler 2000). However, this is not easy to detect. There are several sources of homoplasy other than different branching history, including evolutionary convergence. If several data partitions show different highly discordant trees due to convergence, the only way to see the 'true' tree topology is to combine them. The only weapon a systematist has against convergence is the likelihood that truly independent characters will be subject to different confusing factors and thus the true history may emerge when these independent characters are combined (Barrett *et al.* 1991). Probably all character systems are influenced by constraints that tend to bias phylogeny reconstruction one way or another, yet a combination of very different character data may allow the noise to cancel out, and the historical signal to come through.

Therefore, observing a particular gene or other data partition exhibiting serious conflict with another is not sufficient reason to reject combining them. There must also be additional evidence, outside of the phylogenetic analysis, for reticulation or lineage sorting. The best current examples of such discordance are in shallow analyses, where organellar genomes may have different phylogenies than those of associated nuclear genomes and morphologies (Smith and Sytsma 1990; Rieseberg and Soltis 1991). Barring that sort of clearly explainable discordance via reticulation, all appropriate data should be used, especially in deep analyses because, as argued earlier, reticulation and lineage sorting are much less likely to be problems in deep analyses, while convergence is likely to be a greater problem. But even if its effects may be negligible in many deeper analyses, the problem of reticulation is a difficult one, worthy of a more detailed look.

4.10 Reticulation

As introduced earlier, the tree of life is essentially composed of nested sets of lineages. Look closely at one lineage, and it turns out to be composed of smaller lineages, all the way down to within the organism (e.g. cell lineages and gene genealogies). None of the levels of nested lineages can be considered fundamental (Mishler and Theriot 2000a, b,c)—it depends on the scale of the specific question being asked. To build the large-scale framework of the tree of life one can probably ignore the fine-scale lineages within organisms and between organisms within populations. But to study microevolutionary differentiation processes and design conservation plans at the population level, one needs to look at the fine-scale lineages, and to look at the spread of cancer cells in a body, one needs to look at finer levels still.

The major problem that arises is that these nested sets of lineages are not always proper subsets. Especially at the finer levels, sublineages of a larger lineage may not all have the same history, and/or may not have the same history as the larger lineage. For example, parts of the genome within one organism can have different

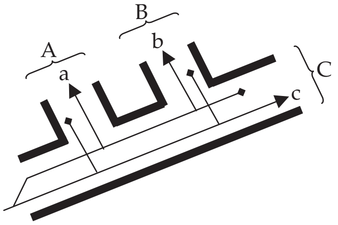

Figure 4.3 Illustration of lineage sorting. Three larger-scale lineages are outlined with dark lines and labeled with capital letters. Three extant smaller-scale lineages are included, together with extinct relatives, and shown with lighter lines and lower-case letters. Note that the relationships of the larger-scale lineages are A(B,C) while the relationships of the smaller-scale lineages are (a,b)c because of the particular pattern of extinction that occurred. This would result in apparent homoplasy at the level of the larger-scale lineages.

histories, for two main reasons. The first of these is *lineage sorting* (see Fig. 4.3), which occurs when genes exist in families within the genome due to past duplication events, and differential extinction has taken place in derived higher-level lineages such that the relationships of the genes appear not to match the relationships of the higher-level lineages (Avise 1989). The problem in this case is one of mistaken homology—paralogy is confused with orthology because not all the gene lineages are present in all higher-level lineages.

The second major reason for differential histories is *reticulation,* which occurs when once separate lineages blend back together. At the genome level, recombination can bring genes with different histories together into a single lineage. Of all the different sources of homoplasy, such as adaptive convergence, gene conversion, developmental constraints, mistaken coding, lineage sorting, and reticulation, the last is the most problematical. This is because reticulation violates a fundamental assumption underlying cladistic analysis, that of a branching model of history. The other factors are all cases of mistaken hypotheses of homology of one sort or another, whereas 'homoplastic' character distributions due to reticulate evolution involve true homologies whose mode of transmission was not tree-like. The possibility of reticulation further complicates the relationship between TUs and characters discussed earlier, since it ensures that some lineages nested inside of larger ones truly have different histories than others.

Because of this important violation of a fundamental cladistic assumption, Hennig (1966) and later Nixon and Wheeler (1990) were correct in focusing on reticulation and the problems it causes for cladistics. However, the problems posed by reticulation are more complicated than their proposed 'solution,' i.e. their suggestion that the species level can be used as a dividing line by supposing that reticulation only occurs below the species level. This assumption (made by many, but not all, cladists) of an abrupt cessation of interbreeding at the species level, separating rampant reticulation below from clean divergent evolution above, was wrong in two important respects. One is the implication that reticulation can be disregarded at higher levels, and the other is the implication that cladistic methods are not appropriate below the species level. Mishler and Theriot (2000a, b, c) refuted both implications; here are their arguments in summary:

(1) There is no consistent demarcation between reticulate and branching relationships at any particular level. Hybridization takes place between clades of various patristic/cladistic degrees of relatedness. Reticulate relationships range from intense (in panmictic, sexually reproducing groups where individual relationships are exclusively reticulate), to less intense (in spatially or temporally subdivided groups where both reticulate and divergent relationships exist among individuals), to none in clonally reproducing organisms. Rare, high-level hybridizations may occur among very divergent lineages, such as among genera of orchids; viral-mediated lateral transfer of genetic material is suspected at much higher levels.

(2) Just as barriers to reticulation are often not complete, reticulation is not a complete barrier to cladistic analysis. There is much phylogenetic structure within named species; indeed, a whole new field of phylogeography was founded to explore this (Avise 1989). We can reconstruct relationships in the face of *some* amount of reticulation (how much is not yet clear, but is amenable to study). For example, McDade (1992) showed that incorporating a few known hybrids in an analysis of 'good' species does *not* seriously affect the cladistic topology of the good species. There may be a self-correcting mechanism here as there

is with other sources of homoplasy: even major convergence (e.g. among cave animals) can be uncovered via cladistic analysis. As with convergence, where the application of cladistic analysis provides the only rigorous basis we have for identifying homoplasy and thus demonstrating non-parsimonious evolution, the only way we can identify reticulation on the basis of character analysis alone is through the application of cladistic parsimony, followed by the examination of homoplasy to attempt to discover its source (see discussion by Vrana and Wheeler 1992). As was pioneered by Slatkin and Maddison (1989), cladistic analysis of non-recombining genes can even be used to measure gene flow between populations. Thus, cladistic analysis can be used to study reticulation, at any level.

(3) Thus, just as there may be no largest cladistic unit for which reticulation is impossible, there may be no smallest 'irreducible' cladistic unit within which no further diverging phylogenetic patterns occur. Ontologically speaking, we are dealing with a fractal pattern again; if you look inside one lineage you see a pattern of divergence of lineages within (and some reticulation, perhaps increasingly greater as one looks at less-inclusive lineages). This fractal pattern of reticulation and branching presents a problem for simple phylogenetic inference. But, as argued above, phenomena such as lineage sorting and reticulation can be discovered as incongruence between organismal and gene phylogenies, or incongruence between different genes or different regions of the genome.

4.11 TUs, characters, and database design

One of the big challenges in modern biology is informatics. There are so many data available, and a number of projects are attempting to represent the information in databases. However existing databases (e.g. GenBank or Tropicos) are essentially a flat file with respect to phylogeny. Data are entered with whatever taxon name happens to be attached to them. The only sense of evolutionary relationships is given by a schema of higher-taxon names (say families and phyla) that can be used to group the basic entries. These higher taxa may or may not be monophyletic, and essentially function as static sorting bins for pulling out the basic records—there is no way to access or display emergent properties of data at higher evolutionary levels or to discover finer-scale patterns at lower levels. In other words, databases are not yet sensitive to the fractal nature of phylogenies (with their many hierachically nested levels). As argued above, there are no basic comparable taxa (terminal or otherwise), or characters. Both TUs and characters are defined with respect to a certain level in the phylogeny.

As a new generation of phylogenetic databases are built (in part coordinated by a large NSF ITR grant supporting a national resource in phyloinformatics, Cyber Infrastructure for Phylogenetic Research (CIPRes); see www.phylo.org), there needs to be much more flexibility built in. The main themes of this chapter need to be explored to appropriately present the richness of phylogenetic data to users. Fundamental open questions that need to be addressed for databases include: (1) how can the elements of the data matrix (TUs, characters, and states) as defined and recognized in some particular study be stored and potentially retrieved for use in a future study at a different level? (2) How can heterogeneous data types (e.g. DNA sequences, genomic rearrangements, morphology) be compared/combined? (3) How can data sets and analyses at very different scales be concatenated (e.g. supertree, compartmentalization, or global approaches as discussed earlier)? (4) How can phylogenetic results at these different concatenated scales, where TUs are nested inside larger ones, and character definitions (e.g. alignments) change as you move up and down the scale, be presented to the community in comprehensible and useable ways?

The centerpiece of all future biological databases will need to be phylogenetic classification, a deeply nested hierarchy of named nodes linked to all available structural and functional data at each level dynamically, as new data enter the database. All biological data fall somewhere on the tree of life, which is the one thing that can unify them all. This new approach to biodiversity informatics will take advantage of the richness of the phylogenetic structure of biological data.

4.12 Tree building

This chapter has focused on the first phase of phylogenetic analysis, building the data matrix, rather than the second phase, building a tree from the matrix. Still, a few words on the latter are appropriate. The simplest model for evaluating congruence among characters (different hypotheses of homology) is equally weighted parsimony (Farris 1983), which remains the preferred method for comparing diverse sorts of characters. Each column in a data matrix can be regarded as an independently justified hypothesis about phylogenetic grouping (the criteria for justifying these individual character hypotheses is discussed above), an individual piece of evidence for the existence of a monophyletic group. Parsimony assumes that an apparent homology is more likely to be due to true homology than to homoplasy, unless evidence to the contrary exists, i.e. a plurality of apparent homologies showing a different pattern (Funk and Brooks 1990; Mishler 1994). Parsimony does involve some simplifying assumptions, i.e. that all character-state changes are similar in their probability of change, and thus they can all be equally weighted. This assumption, while robust, can lead to mistaken reconstructions under some extreme circumstances of asymmetric probabilities of change within and among characters, and in such cases simple parsimony can be modified using more complicated models of change by either character and character-state weighting (Albert *et al.* 1992, 1993; Albert and Mishler 1992) or maximum likelihood approaches (Felsenstein 1981; Yang 1994).

Debates will no doubt continue over how complicated an evolutionary model it is prudent to include in an analysis, but it is clear that all the parsimony and maximum likelihood methods, by using individual character data (specific hypotheses of homology), belong to a related Hennigian family of methods. Fortunately, one important empirical observation is that differential weighting and maximum likelihood have little effect on simple parsimony reconstructions. Weighted parsimony and maximum likelihood topologies are almost always a subset of the equally weighted parsimony topologies, especially when applied to data with an appropriate rate of change for the problem at hand (more on this later). Thus, paradoxically, pursuit of well-supported weighting schemes has ended up convincing many of us of the broad applicability and robustness of equally weighted parsimony (Albert *et al.* 1993). Furthermore, all reconstruction methods work best with 'good data', i.e. characters chosen with respect to a particular level of phylogenetic question. It is with more problematic data (e.g. with a limited number of informative characters, a high rate of change, or strong constraints) that results of different methods begin to diverge. Weighting algorithms and maximum likelihood approaches may be able to extend the use of problematic data, but only if the evolutionary parameters that are biasing rates of change are known. As biases become greater, precise knowledge of them becomes ever more important for avoiding spurious reconstructions. Therefore, given the large number of potential characters made available by modern technology, it is desirable to be highly selective about the characters that are used to address any particular phylogenetic question; to the extent possible, the problematic data should be left out (possibly to be used at a different, more appropriate level: see discussion on compartmentalization in Mishler 1994, and elsewhere in this chapter).

What is the relationship between this chapter emphasizing the data matrix, and the general themes of this book on parsimony? Simple. A rigorously produced data matrix has already been evaluated carefully for potential homology of each feature when being assembled. Everything interesting has already been encoded in the matrix; what is needed is a simple transformation of that matrix into a tree without any pretended value added. Straight, evenly weighted parsimony is to be preferred, because it is a robust method (insensitive to variation over a broad range of possible biasing factors) and because it is based on a simple, interpretable, and generally applicable model. More-complicated models for tree building are fundamentally attempts to compensate for marginal data. Given the surfeit of data available these days, it would be wiser to avoid the use of marginal data!

These issues of how to use phylogenetic markers at their appropriate level to reconstruct the extremely fractal tree of life are likely to be one of the major concerns in the theory of phylogenetics in coming years. In the future, my prediction is that more-careful selection of characters for particular questions (i.e. more-careful and rigorous construction of the data matrix) will lead to less emphasis on the need for modifications to equally weighted parsimony. The future of phylogenetic analysis appears to be in careful selection of appropriate characters (discrete, heritable, independent, and with an appropriate rate of change) for use at a carefully defined phylogenetic level.

4.13 Acknowledgements

This chapter has benefited from analyses and collaborations supported by three NSF grants, and I acknowledge my co-principal investigators for their help in understanding these issues: the Deep Gene Research Coordination Network (DEB- 0090227; http://ucjeps.berkeley.edu/bryolab/deepgene/), the Green Tree of Life Project (EF-0228729; http://ucjeps.berkeley.edu/TreeofLife/), and the ITR grant entitled Cyber Infrastructure for Phylogenetic Research (CIPRes; EF-0331494; www.phylo.org/).