Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Biomedical Informatics 39 (2006) 6–17
Journal of Biomedical Informatics
www.elsevier.com/locate/yjbin

Methodological Review

What's in a character?

Rob DeSalle*

Division of Invertebrates and the Molecular Systematics Laboratories, American Museum of Natural History, 79th Street at Central Park West, New York, NY 10024, USA

Received 9 September 2005
Available online 1 December 2005

Abstract

Systematic analyses are included as integral parts of bioinformatic analysis. The use of phenetic and phylogenetic trees in many of the newer areas of biology create a need for bioinformaticists to understand more completely the nuances of systematic analysis. Any description in comparative biology, universally begins with what information to use in the comparative endeavor. Phylogenetic approaches are no different. The diversity of approaches and phylogenetic questions in systematics have sometimes hindered a precise understanding of what primary data should be collected to perform such analyses. In addition, one should always keep in mind that the objective of systematic organization of entities in nature not only strives to organize those entities in an objective, repeatable and operational way, but also to organize the attributes of the entities in a similar hierarchical context. This paper attempts to describe characters as the basis of all comparative analysis, to describe the diverse kinds of primary data that exist today in biology, genomics, and bioinformatics, and to place these kinds of primary data in the context of the established approaches to tree building.

Keywords: Systematics; Bioinformatics; Phylogenetics; Characters; Likelihood; Microarrays; DNA sequences

1. The process of systematic analysis

Several manuals published in the last decade exist, that provide detailed approaches to systematic analysis [1–20]. One of the best descriptions of how systematic analysis works can be found in Wenzel [12]. His scheme for how systematics should proceed is presented in Table 1. In essence Wenzel points out that two thirds of phylogenetic analysis is concerned with characters (''Establish a matrix'' and ''Establish a weighting scheme''), hence it is most important to define what characters are and how they are used and what the units of systematic analysis are.

Entities in systematic analysis can be either living organisms or inanimate objects. Anything that can eventually have characters coded for it is an entity available to systematic analysis. Some examples of inanimate objects in systematic analysis are the Caminicules [21,22]; http://taxonomy.zoology.gla.ac.uk/rdmp1c/teaching/L4/Evolution/Session2/cam.html] introduced as a model system for systematic analysis, architectural styles [23], and the ''bolt'' example [24]; http://www.pbs.org/wgbh/nova/teachers/activities/2905_link.html] found in elementary primers of systematic analysis. Of course, systematic analysis is much more interesting when descent with modification has been at work, but the so-called ''pattern cladist'' approach, in which heirarchical patterns are postulated and tested with no reference to evolutionary processes, is an interesting and valid approach to systematic analysis. The question of whether evolution or descent with modification is a necessary assumption or is required as background knowledge of systematic analysis is beyond the scope of this review: the reader is referred to Brower [25,26] and Kluge [27] as an introduction to this subject. This debate is a consideration in modern tree building when microarrays are examined for heirarchical data structuring. It is very likely that descent with modification is not at work in many of the cell lines used in the majority of micro-array studies, yet trees are used extensively to

* Fax: +1 212 769 5277.
E-mail address: desalle@amnh.org.

1532-0464/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2005.11.002

Table 1
Wenzel's scheme for systematics (with modifications for this publication)

Prologue: Choose a problem (systematic, nomenclatural, functional, taxonomic or evolutionary) that seems to relate to a group with a single common ancestral entity. Sample the entities as densely as is practical, trying to capture the range of variation of the entities.

I. Establish a matrix
  A. Decide character homology from gross variation
    1. Within characters decide state homology
  B. Code characters
    1. Decide on additivity
    2. Decide on dependence
  C. Choose outgroups to provide root (polarity)
    1. Two or more real entities are preferable
    2. Closer is better than more distant

II. Establish a weighting scheme
  A. Differential or uniform
    1. Character state or branch
    2. Static or dynamic
      a. Single function or iterative

III. Calculate shortest Manhattan distance through the matrix
  A. Take consensus of multiple trees
    1. Decide on Adams or Strict
    2. Decide on character optimization flat choice

Logical choice (for instance derive complexity once, lose it often).

represent the results in micro-array studies (see Planet et al. [28]). This discussion points to the idea that careful consideration of the assumptions inherent to any treebuilding program should be considered when data are analyzed in a phylogenetic context.

Before proceeding, three distinctions must be made with respect to the representation of the entities in systematic analysis. These distinctions involve taxa, phenetics, and homology.

1.1. Taxa

Throughout this article I refer to the subjects of systematic analysis as entities or taxa. My reference to entities is because sometimes the subjects of systematic analysis are not always living organisms. For living entities I refer to the subjects of systematic analysis by the entrenched designation of "taxa" or "taxonomic units." Taxonomic units are either "observed" (OTU) or "hypothetical" (HTU). OTU's are found at the terminals of phylogenetic trees. Even fossils are OTU's and are treated as terminals in trees in systematic analysis. A common misconception of the uninitiated in systematic analysis is that fossils represent ancestral taxonomic units. However, the only ancestral taxonomic units in phylogenetic analysis are hypothetical (i.e., HTU's). HTU's which exist at the nodes of a phylogenetic tree represent entities with the ancestral character states of the OTU's and other HTU's that are situated outward from the more basal HTU's (Fig. 1). They are hypothetical because their character states are determined by reconstruction methods, either parsimony based [29–34] or via likelihood analysis [35,36]. Excellent examples of the utility of reconstructed HTU's can be found in the review by Thornton [106], where a description of reconstructing and resurrecting hypothetical ancestral proteins is given.

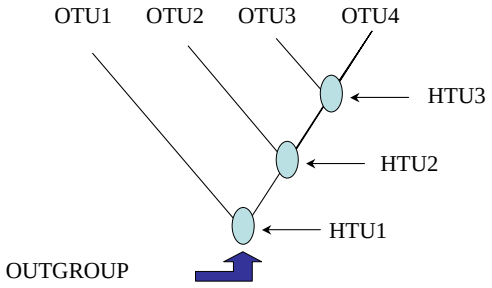

Fig. 1. General diagram of a cladogram showing the terminal positions in the cladogram of Observed Taxonomic Units (OTU's) and the Hypothetical Taxonomic Units (HTU's) that reside at nodes represented by the blue ovals. Observed attribute information characterize the OTU's while the HTU's are inferred using character reconstruction methods. The dark blue arrow at the bottom of the diagram indicates the position of the root usually established by some observed entity that is not considered part of the observed ingroup (OTU's 1 thru 4). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this paper.)

1.2. Phenetics

A distinction can here be made between phenetic clustering diagrams or phenograms and the character based nature of phylogenetic trees. In essence, there are two major kinds of tree building approaches—character based and distance based. Many of the distance based approaches were first described and implemented by the numerical taxonomy school initiated in the late 1950's and early 1960's [37,38] and are grouped under the general heading of phenetic approaches. Phenetic clustering diagrams are often times interpreted as hierarchies, but strictly speaking this is not the case. A phenogram is simply a visual way to present a matrix of distance data and technically does

not represent a hierarchy, because the HTU's at nodes have no attributes. On the other hand, a character based tree does represent a hierarchy in that the HTU's at the internal nodes of the tree have reconstructed attributes and can demonstrate transformation of characters at the internal nodes of a cladogram. Along the same lines it is evident that an HTU in a phenogram is not terribly useful with respect to understanding the changes that have occurred during the divergence of the entities in the analysis. This problem arises from the obliteration of character based information due to the compression of characters into distances in phenetic approaches. Even though an argument can be made that characters can simply be mapped back onto a phenetic clustering diagram, such mapping procedures will often times give imprecise inferences of character change because the phenogram might not be the best representation of divergence of the entities under examination. One highly visible example of this failing is where large amounts of change have occurred in a terminal entity or OTU. Fig. 2 shows such a case where entity B has incurred many changes since its divergence from the common ancestor of entities A and B with respect to entities C and D. When the morphological characters "Hairs," "Wings," and "Legs" are traced on a parsimony tree constructed from these hypothetical data their interpretations are straightforward. "Hairs" is a change that occurs only in Taxon A. "Wings" is a change that is found in the common ancestor of Taxa A and B, and "Legs" is a change that is

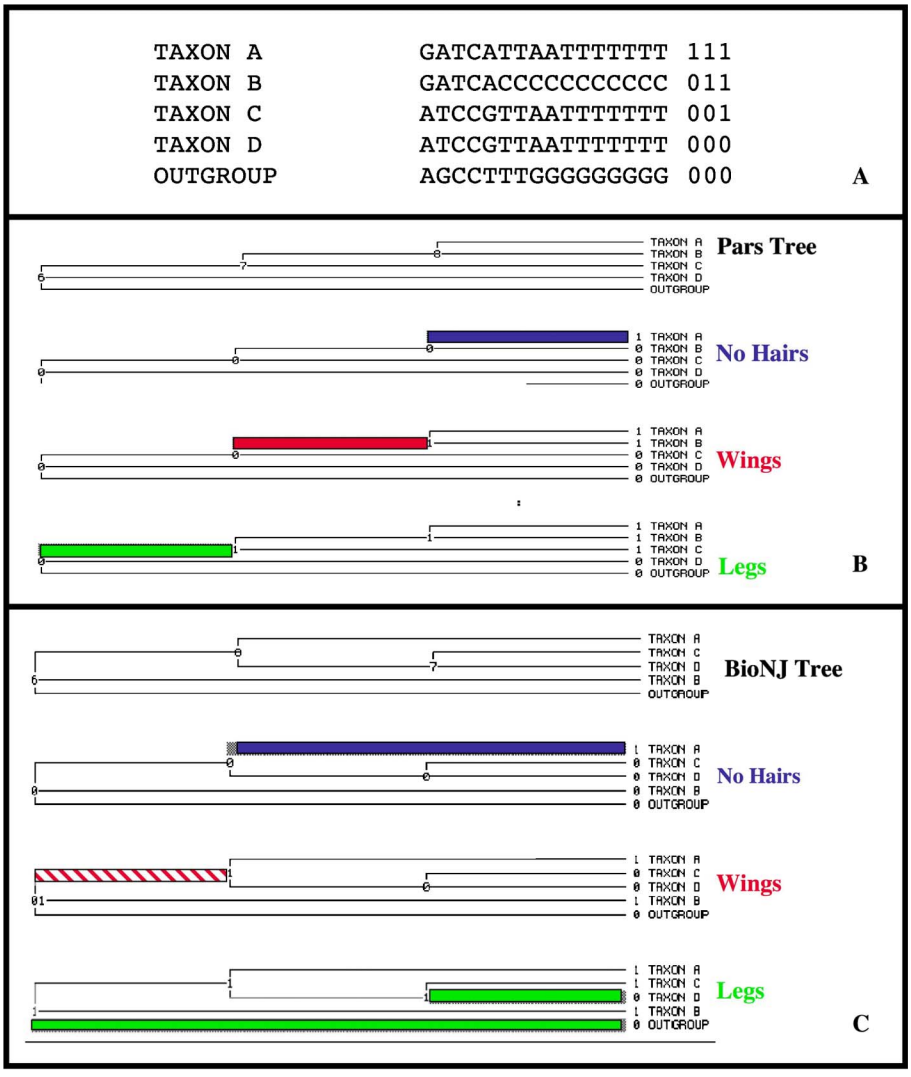

Fig. 2. Hypothetical example showing the difference in character reconstruction based on a maximum parsimony reconstruction and distance analysis of character state data. (A) Hypothetical data set for four ingroup taxa (Taxa A thru D) and an outgroup taxon (OUTGROUP). The first 16 character columns are DNA sequence data for the five taxa. The next three character columns are morphological data for the presence (1) or absence (0) of Hair (blue), Wings (red), and Legs (green) in that order. (B) Maximum parsimony (MP) tree constructed from the matrix in panel A with the three morphological characters mapped onto the tree. Panel C shows the three morphological characters mapped onto a Bio-Neighbor Joining (Bio-NJ; [83]) tree constructed from the data in A. Colored bars on the trees show where characters would be reconstructed for the three characters. A striped bar indicates that the reconstruction is ambiguous. The point of the diagram is that character reconstruction on trees generated using different methods can be different. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this paper.)

Table 2
Some basic terminology

Entity: In the context of this review, an entity is any object (animate or inanimate) that can be placed on the terminal of a phylogenetic or phenetic tree.

Taxon: A classification or group of entities. In the context of the tree of life a taxon would be a group of living organisms in a classificatory scheme (ie, kingdom, phylum, class, order, family, genus, species).

Attribute: A characteristic or property of an individual entity, such as weight, size, or color.

Character: The basic unit of systematic analysis. In the definition of Davis and Nixon [45] a character for a taxon is an attribute that is fixed and different in that taxon from other taxa.

OTU: Observed Taxonomic Unit; an entity or taxon that can be examined for attribute states and hence assessed for characters (see Fig. 1).

HTU: Hypothetical Taxonomic Unit; a hypothetical entity that can be inferred to exist at the nodes of a phylogenetic tree. The characters of HTU's are inferred using parsimony or likelihood methods (see Fig. 1).

Homology: Two characters are homologous when they share topological similarity AND are derived from a most recent common ancestor. In the terminology of de Pinna [61] primary homology represents the recognition of topological similarity of a character and secondary or true homology represents the determination that the character is shared and derived for two or more taxa.

Data sieving: The process of removing character information that is either noninformative or irrelevant to the systematic question at hand.

Distance based approaches: Approaches that attempt to infer relationships of entities based on similarity (distance). In some cases the primary data used in distance based approaches are in similarity (distance) format such as in DNA/DNA hybridization studies. In most cases though, the primary data are character state data and these data are then transformed into distances. These approaches are also referred to as phenetic based.

Character based approaches: Approaches that attempt to infer the relationships of entities based on the untransformed character information. There are two major character based approaches in use—maximum parsimony and maximum likelihood.

found in the common ancestor of Taxa A, B, and C (Fig. 2). The phenogram constructed from these hypothetical data would not place entity A and B as each other's closest relatives, and in fact when interesting attributes such as "Wings" and "Legs" are traced onto the phenogram, they are interpreted as having ambiguous evolutionary change or as a convergence, respectively.

1.3. Homology

This review will not delve into the problem of homology in detail as it is reviewed elsewhere in this special issue. A basic definition of homology is given in Table 2. Throughout this review it is assumed that there are two phases to homology assessment. The first consists of establishing topological similarity (primary homology [61]) as a hypothesis of homology. The second phase consists of testing the hypothesis posed by the primary homology assessment (secondary homology [61]). A successful test of primary homology establishes that the character in question is shared and derived in the entities under study.

2. Any entity or attribute will do (if you are careful)

The job of the systematist—whether it be an organismal systematist or a gene family annotator, or a micro-array analyst, or a medical bioinformaticist—is to discover the relationships of the entities under consideration using the attributes of the entities. At the risk of sounding reductionist, entities can be considered "bags" of attributes. These attributes range from easily viewed anatomical attributes, to more technically difficult to obtain DNA and protein sequences, to emergent properties of entities such as the response of organisms to ecological or behavioral challenge or as medical information. And from an extreme reductionist perspective entities can be looked at all the way from subatomic particles [39] to genes and proteins, to higher taxonomic groups such as the three grand super-domains of living organisms on the planet—Bacteria, Archaea, and Eukarya, to even emergent properties such as language [40–42] and medical information [43,44].

Davis and Nixon [45] made the important distinction between attributes and characters. Characters are established as such when an attribute can be shown to be fixed and different in one OTU relative to other OTU's under examination. Attributes therefore are the primary assessment of an OTU's makeup. When an OTU is made up of a single entity with respect to an analysis, attributes of the OTU are equivalent to characters. In this context, for example, proteins produced from genes in a particular gene family analysis have amino acids that are both the attributes and the characters of the proteins. Likewise, if one uses tree building methods to analyze micro-array data to determine the relationship of genes to one another based on their expression, then the expression intensities on the microarray are both attributes and characters of the various genes spotted on the array. Similarly, when attempting to infer the relationships of different cell lines or cell types to one another based on micro-array data, the spot intensities are both attributes and characters of the cell lines.

A common misconception in the analysis of individual entities within OTU's is that such individual entities have a demonstrable hierarchy or phylogeny. In sexually reproducing organisms, hierarchy is destroyed by the reticulation of organisms and in the context of organisms within an OTU, any phylogeny of the entities in the OTU has no meaning. Only when a single marker is used will the hierarchy resulting from systematic analysis have

meaning and in these cases the hierarchies produced indicate genealogies of the markers that exist in the OTU. For instance, mtDNA sequences are commonly used in population level analysis of organisms. In this case the mtDNA marker (usually D-Loop or control region sequences) because of its maternal inheritance and lack of recombination will be an excellent tracer of females in a population. Likewise the Y chromosome of many organisms experiences a similar clonal inheritance and is a good tracer of males in a population. Even single gene genealogies of organisms in OTU's can be useful in showing the pattern of descent of alleles in the gene genealogy and hence if the gene is involved in pathology the genealogy could be of importance in interpreting disease. But none of these (mtDNA, Y chromosomes or single nuclear genes) in and of themselves is suitable to establish a "phylogeny" of individual entities within a taxon. This problem makes dealing with single genes or single sources of characters problematic when trying to investigate phylogenetic history of OTU's [46–48]. It is important, therefore, to design a systematic or phylogenetic analysis so that the data subject to analysis is capable of answering the question at hand. The problem just touched upon is the "total evidence" [49], "simultaneous analysis" [31,50–55] or "concatenated analysis" [56,57] problem that has been highly visible in the systematics field for the last fifteen years. The brevity of this review prevents a thorough review of the problem. (but see also Planet and Egan in this special issue).

A final issue concerns choosing attributes for systematic analysis. Living organisms have a limited, but enormous number of physical attributes that can be exploited for systematic analysis. Early systematic studies used, almost entirely, morphological characters for systematic analysis. It might at first seem that morphological attributes are somewhat limited in number, and indeed, average numbers of morphological characters in systematic studies, are usually double digit and occasionally triple digit. However, in most cases the morphological attributes are usually sieved [58,46,55] such that only the attributes that are relevant to the phylogenetic question are scored and used. When DNA sequence attributes were first used in systematics in the 1980's, a similar form of data sieving was implemented. For instance, an examination of the relationships of animal phyla would avoid rapidly evolving information, such as DNA sequences from the mitochondrial D-loop region. On the other hand, an examination of sibling species of *Drosophila* would avoid a source of information that was slow to change over evolutionary time, such as the nuclear 18S ribosomal RNA genes. Whole genome sequencing [56] and high throughput methods [52,53] have allowed for the near elimination or at least the promise of eliminating data sieving in the collection of molecular information in systematics. Novel microscopic methods [59,60] and automated anatomical data collection promise to do the same for morphological information.

2.1. Objectivity, operationality, repeatability

One should prefer an organizing system that is objective, operational, and repeatable. So some attributes of entities are better or more appropriate for organizing than others. For instance, the first letter of the name of an entity is an attribute of the organism. One could simply systematize entities alphabetically and be done with it and this would be a highly operational and repeatable mechanism for organizing entities and their attributes. However as we all know, the first letter of the name of an entity most likely does not carry sufficient biological or systematic information to make a meaningful objective statement about the entities under consideration, because of the subjective nature of first letters of names. On the other hand, one of the grand organizing principles of modern biology concerns modification with descent and the history of such modification through common ancestral entities. This approach is both objective and repeatable, and depending on the number of entities is also highly operational.

3. Discreteness

When one examines an entity, there are many ways to characterize the observations. What one must start with is a description of the attribute and some primary hypothesis of homology [61–63]; see also Phillips in this special issue] of the attribute in a wide range of entities. By establishing some primary hypothesis of homology, the various states of the attributes can be established. There are many ways to describe the attribute states of entities [7,64–66]. Some attributes can simply be described with respect to state as being present or absent. Such attributes are called binary discrete attributes. Other examples of discrete attributes in modern molecular biology are the identity of nucleotides in DNA sequences with the four attribute states of G,A,T, and C or the attributes of amino acid residues in proteins with the discrete attribute states of A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W and Y. These kinds of attributes are called multi-state discrete attributes.

Discrete attributes can be ordered. In the case of binary discrete attributes, one can make the assumption that it is more likely to go from one state to the other and in subsequent phylogenetic analysis restrict certain attribute state transitions accordingly called Dollo parsimony. This kind of parsimony makes the assumption that regaining a complex attribute state ($1 \rightarrow 0 \rightarrow 1$) is not allowed, and only single $0 \rightarrow 1$ and multiple $1 \rightarrow 0$ state changes are allowed. In essence, this method gives greater weight to single $0 \rightarrow 1$ changes in a lineage and to any $1 \rightarrow 0$ changes than the multiple $1 \rightarrow 0 \rightarrow 1$ kinds of changes.

Multi-state discrete attributes can similarly be ordered. For instance, a morphological attribute with three states—absent, small, and large (A, S, and L)—can be forced to obey the following ordering in phylogenetic analysis $A \rightarrow S \rightarrow L$. This ordering amounts to giving $A \rightarrow S$ and $S \rightarrow L$ changes greater weight than $A \rightarrow L$ changes.

4. Some attributes are better than others? Formal weighting

4.1. Weighting states of an attribute

A part of characterizing attributes also concerns the quality of the attribute states. For instance, if from background knowledge, one has a good idea that certain state changes are impossible or highly unlikely it would be unwise in a phylogenetic analysis to allow those kinds of attribute state changes to occur at any high frequency in a phylogenetic analysis [67]. It stands to reason that background knowledge might also inform one that certain kinds of state changes are possible but are much less probable than others. In these cases, certain state changes can be given less weight in phylogenetic analysis leading to the process of attribute state weighting. To accomplish this process, a character transition scheme is needed. The best examples of this approach are the state change matrices for amino acid changes in protein sequences and transversion/transition weighting (or Kimura 2 parameter [68]) for nucleotide sequences (see below). A good example of this approach is in nucleotide based studies, where certain attribute state changes in DNA sequences have been observed to be less likely than others. In addition, the actual physical chemistry of nucleotide base structure indicates that certain kinds of attribute state changes in DNA sequences are more likely than others. These kinds of changes are the so-called transitions versus transversions (Fig. 3). It is well known that transitions—purine (C and T) to purine changes and pyrimidine (A and G) to pyrimidine changes—are more likely than transversions—purine to pyrimidine or pyrimidine to purine changes. In this classic case, one can weight certain attribute state changes that involve transitions lower than attributes that involve transversions in phylogenetic analysis (because transitions occur more frequently, they are more prone to convergence).

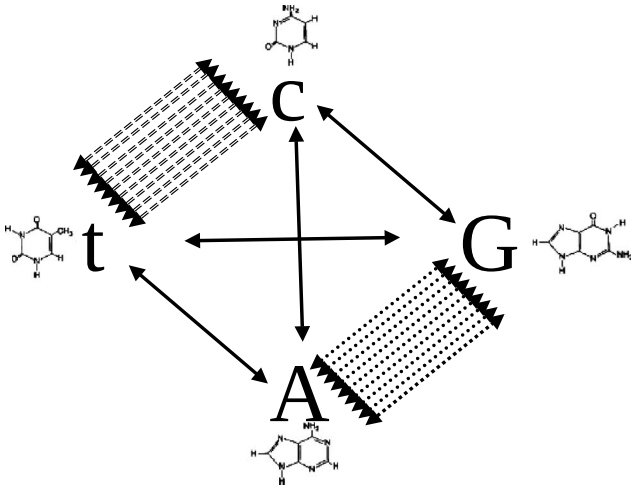

Fig. 3. Diagram showing the general patterns of nucleotide substitution in DNA sequence evolution. G stands for guanine, A for adenine, c for cytosine, and t for thymidine. The upper and lower case letters in these designations indicate the larger size (upper case) of A and G nucleotides that are comprised of two molecular rings (see molecular stick diagram near the G and A) versus the smaller c and t nucleotides (see molecular stick diagram near the c and t). The number of arrows between any two bases indicates the frequency with which substitution occurs between the pairs. The solid arrows represent TRANSVERSIONS (TV) and the multiple dashed arrows indicate TRANSITIONS (TI). The multiple arrows indicate that t to c and A to G transitions are more common than any of the four transversions.

4.2. Weighting attributes relative to one another

Another kind of weighting concerns choosing certain attributes (not attribute states) that appear to be more informative than others from experience or from background knowledge and giving those attributes more weight. In this context, the relative weight of certain classes of attributes (also called process partitions; [69]) has been a subject of intense debate. The most pointed discussion centers around organismal systematics where authors argue for the weighting of morphological or anatomical attributes over molecular. This argument is based on the idea that anatomical attributes are more reliable than molecular data because they tend to converge less frequently than morphological ones. In addition, anatomical attributes can have many more states than nucleotide sequences which have only five (G, A, T, C, and alignment dependent gaps).

4.3. Is not weighting, weighting?

Giving equal weights to all characters is in itself a form of weighting. However, there are strong philosophical reasons for why equal weighting might be superior to differential weighting. Kluge [70] argues that ''precision'' in phylogenetic analysis provides an argument for equal weighting of all characters in a dataset. This argument is related to the explanatory power and testability of hypotheses. When characters are weighted, explanatory power and testability of hypotheses are reduced. This iron-clad argument has not stopped most systematists from using models or a priori reasoning to weight characters and attribute states.

5. Attribute state transformation matrices

5.1. Parsimony

In some cases we are informed about the probability of attribute state changes from the background biology of the entities under examination. For instance, in DNA sequences, while we know transitions occur more frequently than transversions (Fig. 3), we can also attempt to quantify the degree to which transversions are to be favored over transitions. Most phylogenetic analysis packages allow for a character state transformation matrix to influence the phylogenetic analysis. The following attribute state transformation matrix is an example of weighting transver-

sions greater than transitions where the numbers in the matrix represent the cost in number of steps such changes would incur in a phylogenetic analysis.

	G	A	T	C
G	0	1	2	2
A	1	0	2	2
T	2	2	0	1
C	2	2	1	0

Two issues relevant to this kind of weighting matrix with respect to nucleotide sequences require discussion. First, the exact values in the matrix can be very taxon specific. For instance, if one wanted to reconstruct the relationships of maternal lineages within a species using mtDNA D-Loop sequences, the apparent frequency of transitions to transversions is very high [71,72] and the more frequent transitions would be very informative, leading the systematist to rely heavily on transitions AND transversions to reconstruct the maternal genealogy. As divergence of organisms gets larger and larger, the transitions become less reliable because some of the nucleotide positions become saturated with change and exhibit convergence and the apparent frequency of transitions to transversions becomes lower. In this latter case, one would prefer to use transversions OVER transitions in phylogenetic analysis. These transition to transversion ratios can usually be estimated from the actual sequence data or can be a priori set from knowing the relative times of divergence and using a calibrated ratio [72]. Second, the example leaves out a potential fifth character state, alignment dependent gaps (-), but this character state is more difficult to incorporate into the matrix based on frequency of its occurrence.

Amino acid transformations in proteins can be similarly weighted using a broad array of matrices. The most simple transformation matrix used for this purpose is the genetic identity matrix first proposed based on the genetic code. Other transformation matrices have been used in phylogenetic analyses based on quantification of the kinds and frequency of amino acid transformations in proteins in the database. The first matrix constructed for this purpose was the Dayhoff matrix and more recently developed matrices such as the PAM matrices [73] and the BLOSSUM matrices [74] can also be adapted to weight characters in protein based phylogenetic studies.

5.2. Likelihood

The above discussion describes attribute state transformation matrices for parsimony analysis. Likelihood analysis uses similarly constructed transformation matrices to implement models of evolution for proteins and nucleic acids. For instance, a generalized transition—transversion two parameter transformation matrix for maximum likelihood analysis is shown below.

	G	A	T	C
G	0	α	β	β
A	α	0	β	β
T	β	β	0	α
C	β	β	α	0

Where α is the probability of transitions occurring and β is the probability of a transversion occurring. Fig. 4 shows the various models that have been used in likelihood analysis and their relationship to one another (after [75]). In general, the transformation probabilities in these matrices are calculated from the data themselves and the appropriateness of particular models relative to one another is determined using a hierarchical likelihood ratio test of each model (MODELTEST; [76]).

It also stands to reason that not all positions in a DNA sequence or protein sequence will have equally probable site specific variation. In these cases, probability of site specific heterogeneity is calculated using a γ function whose shape is determined by an α parameter (see Fig. 13 in [75]). It has been suggested that varying the α parameter in likelihood analysis has a much larger affect on accuracy

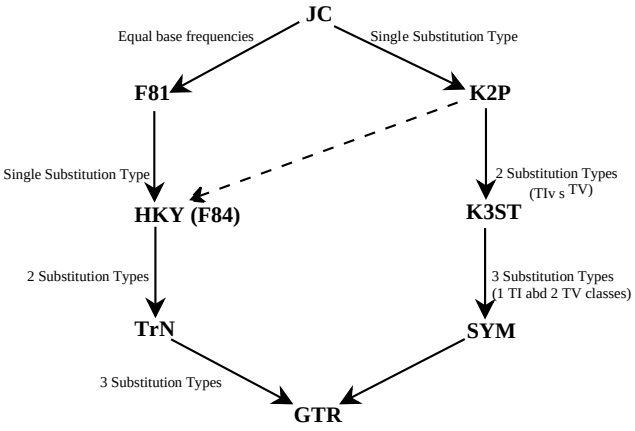

Fig. 4. Hierarchical relationships of the general time reversible family of substitution models used in maximum likelihood searches and in calculating distances for distance methods after [75]. Arrows indicate the class or identity of an assumption of molecular evolution incorporated into a model of nucleotide change that convert the model with the least number of assumption (JC) to the models with more assumptions. Model abbreviations after [75]; JC = Jukes and Cantor [95] the least assumptive model in the maximum likelihood hierarchy; F81 = Felsenstein [96] where the assumption of equal base frequencies is made; K2P = Kimura [97] two parameter model where the assumption of a single substitution type is made; the K2P model can give rise to the HKY (F84) = Hasegawa, Kishino, and Yano model [98] where the more restrictive assumption of two substitution types is made and the K3ST= Kimura [99] three substitution type model where the K2P model has a further assumption of equal base frequencies applied; the F81 model gives rise to the HKY (F84) model; the TrN = Tamura and Nei [100] model is related to the HKY (F84) model such that the further assumption of two substitution types is made; the SYM = model described by Zharkikh, 1994 [101] where three substitution types are assumed to be part of the model; and finally the GTR = Generalized Time Reversal model (Lanave, [102]; Tavare [103], Rodriguez et al [104]) model is the most assumption laden with the possibility of up to six base change types. See also Fig. 10.2 in [105, p. 263].

of the analysis than varying the transformation matrices [77]. The α parameter is often times set at 0.5 to reflect a γ distribution where large numbers of sites do not vary and a small number of sites contain most of the sequence variability. Another way to set the α parameter in likelihood analysis is to calculate α directly from the data matrix being analyzed.

Some controversy exists over the treatment of attributes in the context of preference for either method. The discussion of the relative merits of parsimony and likelihood (as well as Bayesian methods) is both philosophical (for a recent review see [78]) and theoretical. In an important theoretical study Tuffley and Steele [79] demonstrated that the likelihood approach converges on parsimony "if it is assumed that the characters do not evolve by a process common to all the characters" [12, p. 27]. In essence, what is being stated is that likelihood approaches specify what the process is for the characters used in an analysis and that parsimony is a more robust approach because it more closely approximates estimating the exact causes of a character state. Wenzel [12] further suggests that parsimony is a more valid approach when no common process exists for all characters. This suggestion does not assert that there should be all different processes just as Farris [80] first pointed out. Because molecular and morphological characters appear to behave as if there are no common processes for all characters it would follow that parsimony is a more appropriate approach to phylogenetic analysis.

6. Continuous attributes

6.1. Transforming discrete attribute data to continuous

Phenetic approaches can start with distance or similarity measures of entities or can start with discretely distributed attribute information (such as molecular sequence information) that are then transformed into distance or similarity matrices. A more modern version of phenetics is the widely used Neighbor Joining method (NJ; [81–84]), which can routinely take discrete DNA sequence, protein sequence data and transform these kinds of sequence data into a matrix of distances that can then be used in the NJ algorithm. The transformation of the molecular sequence attributes into distances can be directed by any number of models of DNA or protein sequence evolution. In addition, the NJ method for nucleotide or protein sequences is extremely rapid compared to parsimony, likelihood, and Bayesian approaches.

6.2. "Truly" continuous attributes

Attributes of organisms that require continuous numerical description such as measurements like the length of an entity or one of its structures or the counts of anatomically repeated structures like hairs or scales, are good examples of continuous attributes that have been used in systematics. Discrete or discontinuous attributes are those with state values that do not overlap with other state values (Fig. 5). Classic examples of continuous data in molecular based data for systematics are immuno-diffusion data and DNA/DNA hybridization data. A more modern example of continuous data in biology are data derived from micro-array spot intensities. Fig. 5 shows the distinction between continuous and discrete attributes.

A	B	C	D
continuous not gapped	discrete	discrete	discrete
0.1	0.1	G	absent
0.2	0.1	G	absent
0.3	0.1	G	absent
0.4	0.5	T	absent
0.5	0.5	T	absent
0.6	0.5	T	present
0.7	0.8	C	present
0.8	0.8	C	present
0.9	0.8	C	present
1.0	0.8	C	present

Fig. 5. Diagram showing the distinction between discrete versus continuous attribute information. Column A shows a set of 10 measurements that fall into a continuum with 0.1 increments. Column B shows numerical measurements with three distinct clusters of measurements or discrete character states oi 0.1, 0.5, and 0.8. Column C shows DNA sequence information where the bases themselves are discrete character states. Column D shows discrete attribute states present and absent for a morphological structure.

The classical examples of immuno-diffusion data and DNA/DNA hybridization data are immediately amenable to phenetic analysis, but are immune to character based analysis such as parsimony (see [55]). The more modern example of continuous micro-array spot intensities is unique in that thousands of continuous intensity values are generated. There are many approaches to using continuous micro-array data for tree building. In the phenetic approaches to analyzing micro-array data, the thousands of spot intensity measures for two entities are transformed into a single similarity measure for those two entities. Many of these are based on simple phenetic approaches, and others use multivariate statistical analysis such as centroid analysis [85]. Continuous micro-array data can also be "discretized" using "binning" methods [86,87]. Fig. 6 shows some hypothetical examples of binning and details the problems with the various approaches. Binning can result in the discretizing of values into different bins that are more similar to each other than they are to values in their own bins, and the converse problem of placing two values in the same bin that are more dissimilar to each other than values in other bins. Sarkar et al. [88] and Planet et al. [28] have suggested that the binning process to discretizing continuous micro-array data should score as missing values that are in between the "greatest" upper and "smallest" lower values for each spot on a microarray. In this way, the values that are problematic with respect to the discussion below are scored as missing. One might

A	B	C	D	E
Raw	BINNING **rounding**	BINNING 2 bin size = 0.2 **0.1 - 0.4 = 0 / 0.7 - 1.0 = 1**	BINNING 3 bin size = 0.6 **0.1 - 0.2 = 0 / 0.9 - 1.0 = 1**	BINNINJG 4 bin size = 0.1 **0.1 - 0.3 = 0 / 0.5 - 0.7 = 1.5 / 0.9 - 1.0 = 1**
0.1	0	0	0	0
0.2	0	0	0	0
0.3	0	0	?	0
0.4	0	0	?	?
0.5	0	?	?	0.5
0.6	1	?	?	0.5
0.7	1	1	?	0.5
0.8	1	1	?	?
0.9	1	1	1	1
1.0	1	1	1	1

Fig. 6. Binning continuous attribute information to discretize the attribute states. Column A shows the raw continuous attribute information; column B shows the result of using a binning strategy of rounding. Column C shows the results of binning strategy where the internal bin size (the range of measurements scored as missing) is 0.2. Column D shows a binning strategy where the internal missing bin is 0.6 units large. Column E shows a binning strategy where there are two internal bins each 0.1 units large separating three attribute states 0, 0.5, and 1.0.

argue that discarding data in this way are a poor way to proceed, but there are two aspects about the nature of micro-array data that might warrant treatment as such. First, a researcher might be interested only in those genes or cell lines where the array spot is turned ''on'' intensely or ''off'' intensely, thus rendering those spots with intermediate intensity as irrelevant to significant gene interactions. Second, because there are thousands of spots on a micro-array, discarding information is not as problematic for obtaining a result as one might think. In fact, in some array studies examined by Sarkar et al. [88], Nearly 90% of the information in a micro-array data set are scored as missing using their binning method, yet the inferences obtained after such binning are strong even with such a great proportion of the data scored as missing.

7. Recent advances in character interpretation

Several advances in the way characters are viewed have been made in the last five years. All of these advances involve interpreting attribute information in more complex manners. One of the most innovative approaches treats attribute information such as DNA sequences as a problem of optimization on a phylogenetic tree. In this way, the treatment of sequence data is transformed away from alignment and more toward optimizing large strings of sequences. The so-called direct optimization method results in a phylogenetic tree with optimized strings of characters at each node in the tree without an overall alignment as a starting point [89]. This approach is implemented in the POY [90] software package.

Another important method implementing strings of characters comes from Albert et al., [91,92] and Albert [17]. This approach was developed to allay the fears of some systematists that sequence data were simply too inconsistent because of back mutation and subsequent convergence. The solution to this problem comes from the earliest work on restriction sites where several researchers recognized that it is easier to lose a specific restriction site in parallel than to gain the same restriction site in parallel. Hence the solution is to score characters as strings of DNA sequence information. Albert [17, p. 9] describes the strategy to overcome this problem of inconsistency in rapidly evolving DNA sequences as follows: ''If one were to code only those completely matching strings beginning at certain nucleotide positions, especially larger and larger ones, then these should be rather conservative characters for deep branchings within phylogenetic problems.'' Albert furthermore recognized that DNA sequences need not be the only kind of attribute information that can be coded as such and mentions that amino acid data, intron/exon data and gene presence absence data at the level of the genome or even gene order information. Farris and Kallersjo (Hennig Meetings, Gottingen, Germany,1999) have also presented a related method called r-states or the supersites approach where ''strings of nucleotides are recognized beginning at nucleotide 'W' and then parsed downwards through the matrix, recognizing as many character states as necessary to account for differences within the strings'' [17, p. 8]. Such an approach generates a much larger character state space than with simple single nucleotide or amino acid states.

One final new method in understanding characters was developed specifically to handle micro-array information once the spot intensity data have been ''binned'' (see above). This method called character attribute organization system (CAOS; [28,88,93]) is an extension of the population aggregation analysis (PAA; [45]) approach. Fig. 7 shows the basis for the approach which allows for diagnostics to be discovered in aggregates of entities. The approach

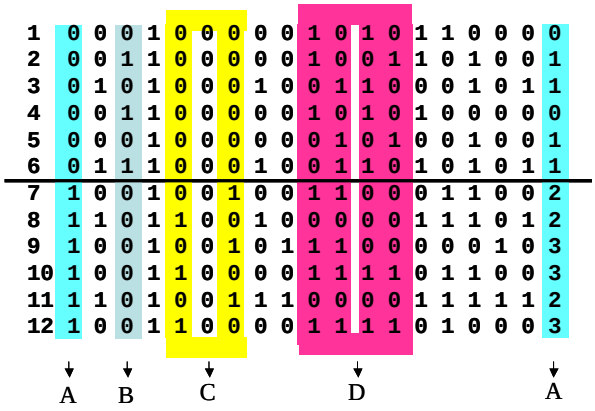

Fig. 7. Hypothetical example of Character Aggregate Organization System (CAOS; Sarkar et al., [80]) in action. The 12 series of 0, 1, 2, and 3 represent discrete scores for 19 attributes for two populations of six individuals. The solid horizontal line through the middle of the matrix represents a geographical barrier between the two populations. (A) 0/1/2/3/ attributes in these columns are purely diagnostic characters (sensu Davis and Nixon, [45]). (B) 0/1 attributes in this column are not purely diagnostic, but rather the 1's in the three individuals in the top population are "private" to that population. (C) 0/1 attributes in the two columns by themselves constitute two private attributes. However in combination these two columns provide a "pure" diagnostic combination (00 versus 01 or 10; this kind of attribute is called a"compound pure" in the terminology of Sarkar et al., [80]). (D) The four columns marked by the shading for D are neither diagnostic nor private. Yet in combination the four columns provide a diagnostic system for the top population versus the bottom. The top population is diagnosed by 10 01 or 01 10 or 10 10 or 01 01 while the bottom population is diagnosed by 00 00 or 11 11 or 11 00 oe 00 11.

extends Davis and Nixon's [45] original approach to diagnosis by including information from "unfixed" attributes such as those attributes with "private" allele like distributions in populations. By combining two or more attributes together with "private" distributions in one entity, diagnostics can be discovered. Sarkar et al. [85] also point out that strictly polymorphic attributes can in combination become diagnostic (see Fig. 7). The CAOS approach has already been used to examine several cancer micro-array studies and has also been suggested as an important analytical approach for "DNA barcoding" [94] and for annotation of genomes [88].

8. Summary

The revolution in modern biology and medicine we are currently witnessing has resulted in a need for more precise analytical approaches. Indeed the last decade has seen a proliferation of model based approaches to understanding the attributes of entities such as in likelihood modeling. Other parsimony based approaches, such as superstrings or supersites analysis as well as direct optimization analysis of DNA sequences have pushed this field further too. Character optimization on well corroborated hierarchical arrangements of entities is an area of modern systematics that has also advanced in the last few years. The application of these approaches to more bioinformatically oriented subjects such as medical informatics, gene ontology, and genomics has also been demonstrated by several researchers and the future will most likely see systematic approaches infiltrate many other areas of modern bioinformatics.

Careful consideration of the goals of any analysis needs to be made before applying an approach. If one wishes to obtain hierarchical interpretations of information from these new sources of bioinformatics, then a more complete understanding of the nature of the entities and attribute information of those entities needs to be accomplished. This review attempts to place attribute information in a character based context and also suggests that character based approaches are the most amenable and philosophically appropriate methods for understanding hierarchy of entities. However if hierarchy is not the goal of an analysis, then other methods may be more appropriate.

References

[1] Zimmer E, White T, Cann R, Wilson A, editors. Methods in molecular evolution: producing the biochemical data, methods in enzymology; 1993. 224, p. 725.
[2] Hillis DM, Moritz C, Mable BK. Molecular systematics. 2nd ed. Sunderland, Massachusetts: Sinauer Press; 1996.
[3] Ferraris JD, Palumbi SR. Molecular zoology advances, strategies, and protocols. New York: Wiley Press; 1996.
[4] Page RDM, Holmes EC. Molecular evolution: a phylogenetic approach. Oxford: Blackwell Scientific; 1998.
[5] Nei M, Kumar S. Molecular evolution and phylogenetics. Oxford: Oxford University Press; 2000.
[6] Scotland RW, Pennington RT. Homology and systematics: coding characters for phylogenetic analysis. London: Chapman & Hall; 2000.
[7] Schuh RT. Biological systematics: principles and applications. Ithaca, New York: Cornell University Press; 2000.
[8] Wiens JJ. Phylogenetic analysis of morphological data. Washington: Smithsonian Institution Press; 2000.
[9] Hall BG. Phylogenetic trees made easy a how-to manual for molecular biologists. Sunderland, Massachusetts: Sinauer Associates; 2001.
[10] DeSalle R, Giribet G, Wheeler W. Methods and tools in biosciences and medicine: techniques in molecular evolution and systematics. New York: Birkhauser-Verlag; 2002.
[11] DeSalle R, Giribet G, Wheeler W. Molecular systematics and evolution: theory and practice. New York: Birkhauser-Verlag; 2002.
[12] Wenzel JW. Phylogenetic analysis: the basic method. In: DeSalle R, Giribet G, Wheeler W, editors. Techniques in molecular systematics. Basel: Birkhäuser; 2002. p. 4–30.
[13] Judd WS, Campbell CS, Kellogg EA, Stevens PF, Donoghue MJ. Plant systematics: a phylogenetic approach. 2nd ed. Sunderland, Massachusetts: Sinauer Associates; 2002.
[14] Page RDM. Tangled trees: phylogeny, cospeciation and coevolution. IL: University of Chicago Press; 2002.
[15] Cracraft J, Donoghue MJ. Assembling the tree of life. New York: Oxford University Press; 2004.
[16] Avise JC. Molecular markers, natural history, and evolution. 2nd ed. Sunderland, Massachusetts: Sinauer Associates; 2004.
[17] Albert VA. Parsimony, phylogeny, and genomics. London: Oxford University Press; 2005.
[18] Zimmer EA, Roalson E. Methods in enzymology 395: producing the biochemical data B. Burlington: Elsevier; 2005.
[19] Salemi M, Vandamme A-M. A practical approach to DNA and protein phylogeny. Cambridge: Cambidge Press; 2003.

[20] Felsenstein J. Inferring phylogenies. Sunderland MA: Sinauer; 2004.
[21] Sokal RR. A phylogenetic analysis of the caminalcules. I. The data base. Syst Zool 1983;32:159–84.
[22] Gendron RP. The classification & evolution of caminalcules. Am Biol Teach 2000;62:570–6.
[23] Hersey G. The monumental impulse: architecture's biological roots. Cambridge: MIT Press; 1999.
[24] Anderson OR, Randle D, Covotsos T. The role of ideational networks in laboratory inquiry learning and knowledge of evolution among seventh grade students. Sci Edu 2001;85:410–25.
[25] Brower AVZ. Evolution is not a necessary assumption of cladistics. Cladistics 2000;16:143–54.
[26] Brower AVZ. Homology, and the inference of systematic relationships: some historical and philosophical perspectives. In: Scotland RW, Pennington RT, editors. Homology and systematics: coding characters for phylogenetic analysis. London: Chapman & Hall; 2000. p. 10–21.
[27] Kluge AG. Parsimony with and without scientific justification. Cladistics 2001;17:199–210.
[28] Planet PJ, DeSalle R, Sidall ME, Bael T, Sarkar IN, Stanley SE. Systematic analysis of DNA microarray data: ordering and interpreting patterns of gene expression. Genome Res 2001;11:1149–55.
[29] Maddison WP. Squared-change parsimony reconstructions of ancestral states for continuous-valued characters on a phylogenetic tree. Syst Zool 1991;40:304–14.
[30] Maddison WP. Calculating the probability distributions of ancestral states reconstructed by parsimony on phylogenetic trees. Syst Biol 1995;44:474–81.
[31] Nixon KC, Carpenter JM. On simultaneous analysis. Cladistics 1996;12:221–41.
[32] Nixon KH, 2000. *WINCLADA* (L. H. Bailey Hortorium, Cornell University, Ithaca, New York (kcn2@cornell.edu)) version 0.9.99m24.
[33] Swofford DL, 2002. PAUP*. Phylogenetic analysis using parsimony (* and other methods), version 4.0. Sinauer Associates, Sunderland, Massachusetts.
[34] Maddison DR, Maddison WP. MacClade version 4: analysis of phylogeny and character evolution. Sunderland Massachusetts: Sinauer Associates; 2000.
[35] Pagel M. The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. Syst Biol 1999;48:612–22.
[36] Cunningham CW, Omland K, Oakley T. Reconstructing ancestral states, a critical reappraisal. Trends Ecol Evol 1998;13:361–8.
[37] Sneath PH, Sokal RR. numerical taxonomy—the principles and practice of numerical classification. San Francisco: W.H. Freeman; 1973.
[38] Sokal RR, Sneath PHA. Principles of numerical taxonomy. San Francisco: W.H. Freeman and Company; 1963.
[39] Gell-Mann M, 2005. Eightfold way. Encyclopedia Britannica. Encyclopedia Britannica Premium Service. 22 August 2005 http://www.britannica.com/eb/article-9032138.
[40] Gray RD, Jordan FM. Language trees support the express-train sequence of Austronesian expansion. Nature 2000;405:1052–5.
[41] McMahon A, McMahon R. Finding families: quantitative methods in language classification. Trans Philol Soc 2003;101:7–55.
[42] Rexova K, Frynta D, Zrzavy J. Cladistic analysis of languages: Indo-European classification based on lexicostatistical data. Cladistics 2003;19:120–7.
[43] Sarkar IN, Planet PJ, DeSalle R, Figurski DH. Knowledge acquisition of organized sets (KAOS) of clinical data. AMIA Annual Meeting: Washington, DC; 2001a.
[44] Sarkar IN, Planet PJ, DeSalle R, Figurski DH. Knowledge aggregation and separation through a novel classification algorithm. AAAS Annual Meeting and Innovation Exposition: 2001b.
[45] Davis JI, Nixon KC. Populations, genetic variation, and the delimitation of phylogenetic species. Syst Biol 1992;41:421–35.
[46] Brower AVZ, DeSalle R. Practical and theoretical considerations for choice of a DNA sequence region in insect molecular systematics, with a short review of published studies using nuclear gene regions. Ann Entomol Soc Am 1994;87:702–16.
[47] DeSalle R, Brower AVZ. Process partitions, congruence and the independence of characters: inferring relationships among closely-related Hawaiian *Drosophila* from multiple gene regions. Syst Biol 1997;46:751–64.
[48] Miyamoto MM, Fitch WM. Testing species phylogenies and phylogenetic methods with congruence. Syst Biol 1995;44:127–37.
[49] Kluge AG. A concern for evidence and a phylogenetic hypothesis for relationships among Epicrates (Boidae, Serpentes). Syst Zool 1989;38:1–25.
[50] Baker RH, DeSalle R. Multiple sources of character information and the phylogeny of Hawaiian drosophilids. Syst Biol 1997;46:654–73.
[51] Gatesy J, Milinkovitch M, Waddell V, Stanhope M. Stability of cladistic relationships between Cetacea and higher-level artiodactyl taxa. Syst Biol 1999;48:6–20.
[52] Gatesy J, Matthee C, DeSalle R, Hayashi C. Resolution of a supertree/supermatrix paradox. Syst Biol 2002;51:652–64.
[53] Gatesy J, Amato G, Norell M, DeSalle R, Hayashi C. Combined support for wholesale taxic atavism in gavialine crocodylians. Syst Biol 2003;52:403–22.
[54] de Queiroz A, Donoghue MJ, Kim J. Separate versus combined analysis of phylogenetic evidence. Annu Rev Ecol Syst 1995;26:657–81.
[55] Brower AVZ, DeSalle R, Vogler AP. Gene trees, species trees, and systematics: a cladistic perspective. Annu Rev Ecol Syst 1996;27:423–50.
[56] Rokas A, Williams B, et al. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 2003;425(6960):798–804.
[57] Gatesy J, Baker RH. Hidden likelihood support in genomic data: can forty-five wrongs make a right? Syst Biol 2005;54:483–92.
[58] Novacek MJ. Fossils as critical data for phylogeny. In: Novacek MJ, Wheeler QD, editors. Extinction and phylogeny. New York: Columbia University Press; 1992. p. 46–88.
[59] Klaus AV, Kulasekera VL, Schawaroch V. Three-dimensional visualization of insect morphology using confocal laser scanning microscopy. J Microsc 2003;212:107–21.
[60] DeSalle R, Agosti D, Whiting M, Perez-Sweeney B, Remsen J, Baker R, et al. Crossroads, milestones and landmarks in insect development and evolution: implications for systematics. Aliso 1996;14:305–21.
[61] de Pinna MCC. Concepts and tests of homology in the cladistic paradigm. Cladistics 1991;7:367–94.
[62] Brower AVZ, Schawaroch VA. Three steps of homology assessment. Cladistics 1996;12:265–72.
[63] Hall B. Homology: the hierarchical basis of comparative biology. San Diego: Academic Press; 1994.
[64] Stevens PF. Character states, morphological variation, and phylogenetic analysis: a review. Syst Bot 1991;15:553–83.
[65] Pleijel F. On character coding for phylogeny reconstruction. Cladistics 1995;11:309–15.
[66] Rae TC. The logical basis for the use of continuous characters in phylogenetic systematics. Cladistics 1998;14:221–8.
[67] Neff N. A rational basis for a priori character weighting. Syst Zool 1986;35:110–23.
[68] Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 1980;16:111–20.
[69] Bull JJ, Huelsenbeck JP, Cunningham CW, Swofford DL, Wadell PJ. Partitioning and combining data in phylogenetic analysis. Syst Biol 1993;42:384–97.
[70] Kluge AG. The science of phylogenetic systematics: explanation, prediction, and test. Cladistics 1999;15:429–36.

[71] Brown W, George M, Wilson AC. Rapid evolution of animal mitochondrial DNA. Proc Natl Acad Sci USA 1979;76:1967–71.
[72] DeSalle R, Freedman T, Prager EM, Wilson AC. Tempo and mode of sequence evolution in mitochondrial DNA of Hawaiian *Drosophila*. J Mol Evol 1987;26:157–64.
[73] Altschul SF. Amino acid substitution matrices from an information theoretic perspective. J Mol Biol 1991;219:555–65.
[74] Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 1992;89:10915–9.
[75] Swofford DL, Olsen GJ, Waddell PJ, Hillis DM. Phylogenetic inference. In: Hillis DM, Moritz C, Mable B, editors. Molecular systematics. 2nd ed. Sunderland, Massachusetts. San Francisco, California: Sinauer Associates; 1996. p. 407–514.
[76] Posada D, Crandall KA. Modeltest: testing the model of DNA substitution. Bioinformatics 1998;14(9):817–8.
[77] Cunningham CW, Zhu H, Hillis DM. Best-fit maximum likelihood models for phylogenetic inference: empirical tests with known phylogenies. Evolution 1998;52:978–87.
[78] Helfenbein KG, DeSalle R. Falsifications and corroborations: Karl Popper's influence on systematics. Mol Phylogenet Evol 2005;35:271–80.
[79] Tuffley C, Steel MA. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. Bull Math Biol 1979;59(3):581–607.
[80] Farris S. The information content of the phylogenetic system. Syst Zool 1979;28:483–519.
[81] Saitou N, Nei M. The neighbor joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 1987;4(4):406–25.
[82] Ota S, Li WH. Njml: a hybrid algorithm for the neighbor-joining and maximum likelihood methods. Mol Biol Evol 2000;17(9):1401–9.
[83] Gascuel O. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol 1997;14(7):685–95.
[84] Atteson K. The performance of neighbor-joining methods of phylogenetic reconstruction. Algorithmica 1999;25:251–78.
[85] Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, et al. Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature 2000;406:536–40.
[86] Manduchi E, Grant GR, McKenzie SE, Overton GC, Surrey S, Stoeckert Jr CJ. Bioinformatics 2000;16:685–98.
[87] Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M, Yakhini Z. J Comput Biol 2000;7:559–83.
[88] Sarkar IN, Planet PJ, Bael TE, Stanley SE, Siddall M, DeSalle R, et al. Characteristic attributes in cancer microarrays. J Biomed Inform 2002;35:111–22.
[89] Wheeler WC. Fixed character states and the optimization of molecular sequence data. Cladistics 1999;15:379–85.
[90] Wheeler WC, Gladstein DS, De Laet J. POY. Version 3.0. ftp.amnh.org/pub/molecular/poy (current version 3.0.11). Documentation by Daniel Janies and Ward Wheeler. Commandline documentation by J. De Laet and W.C. Wheeler; 1996–2003.
[91] Albert VA, Williams SE, Chase MW. Carnivorous plants: phylogeny and structural evolution. Science 1992;257:1491–5.
[92] Albert V. Cladistic relationships of the slipper orchids (Cypripedioideae: Orchidaceae) from congruent morphological and molecular data. Lindleyana 1994;9:115–32.
[93] Sarkar IN, Thornton J, Planet PJ, Figurski DH, Schierwater B, DeSalle R. An automated phylogenetic key for classifying homeoboxes. Mol Phylogenet Evol 2002;24:388–99.
[94] DeSalle R, Egan MG, Siddall M. The unholy trinity: taxonomy, species delimitation and DNA barcoding. Philos Trans R Soc Lond Biol Sci 2005;360:1905–16.
[95] Jukes TH, Cantor CR. Evolution of protein molecules. In: Munro HN, editor. Mammalian protein metabolism. New York: Academic Press; 1969. p. 21–123.
[96] Felsenstein J. Evolutionary tree from DNA sequences: a maximum likelihood approach. J Mol Evol 1981;17:368–76.
[97] Kimura M. A simple method for estimating evolutionary of base substitution through comparative studies of nucleotide sequences. J Mol Evol 1980;16:111–20.
[98] Hasegawa M, Kishino H, Yano T. Dating the human-ape split by a molecular clock of mitochondrial DNA. J Mol Evol 1985;22:160–74.
[99] Kimura M. Estimation of evolutionary distances between homologous nucleotide sequences. Proc Natl Acad Sci USA 1981;78:454–8.
[100] Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 1993;10:512–26.
[101] Zharkikh A. Estimation of evolutionary distances between nucleotide sequences. J Mol Evol 1994;39:315–29.
[102] Lanave C, Preparata G, Saccone C, Serio G. A new method for calculating evolutionary substitution rates. J Mol Evol 1984;20:86–93.
[103] Tavare S. Some probabilistic and statistical problems in the analysis of dna sequences. Lect Math Life Sci 1987;17:57–86.
[104] Rodriguez F, Oliver JL, Marin A, Medina JR. The general stochastic model of nucleotide substitution. J Theor Biol 1990;142:485–501.
[105] Posada D. Selecting a model of evolution. In: Salemi M, Vandamme A-M, editors. A practical approach to DNA and protein phylogeny. Cambridge: Cambidge Press; 2003. p. 256–82.
[106] Thornton JW. Resurrecting ancient genes: experimental analysis of extinct molecules. Nat Rev Genet 2004;5:366–75.