Letter to the Editor

The Triangle Inequality and Character Analysis[1]

Ward C. Wheeler
Department of Invertebrates, American Museum of Natural History

Introduction and Background

With the advent of both analytical and evolutionary models that specify complex character-transformation weights (or probabilities) and computer programs in which they are implemented, the use of multistate characters with elaborate state-transformation costs (matrix characters; Sankoff and Rousseau 1975; Sankoff and Cedergren 1983) is increasing. What has not yet appeared, however, is a general discussion of these characters as they are implemented in various analyses. Since the most prominent use of these characters comes in the phylogenetic analysis of molecular data, it is in this area that the most attention is required.

In its simplest incarnation, a matrix character is one which possesses a matrix of values—each cell of which describes a unique cost for a possible character transformation. These costs are the number of steps required to transform one state into another. In a nonadditive or unordered Fitch-type analysis (Fitch 1971; Farris 1988), all transformations are equally costly—each constitutes a single step. Programs such as Swofford's (1990) PAUP allow the definition of elaborate "step matrices" in which the number of steps required by any transformation can be specified. This matrix may be as simple as a binary character with different costs for forward changes and reversals or as complex as a multistate character with scores of specified transformations. As an example, character models such as Dollo parsimony (Farris 1977) and irreversibility (Camin and Sokal 1965) are, in essence, simple matrix characters, specifying asymmetries in transformation between two states. In both of these cases, changes in one direction are more costly than the reverse, yielding a quantitative polarity statement.

Multistate characters allow more complex delineations. In the case of nucleic acid sequence data, there are four nucleotide bases and 12 transformations to be specified, although not all of these are necessarily unique [actually 16, but the identity transformations play no role in parsimony analysis (Wheeler 1990*a*)]. Most commonly, the matrix specifies only two types of transformation—transitions (purine to purine, A↔G; and pyrimidine to pyrimidine, C↔T/U) and transversions (purine to pyrimidine and the reverse, such as A↔T) (Brown et al. 1982; Liu and Beckenbach 1992). Transition-transversion ratios are frequently specified in analyses, regardless of whether the chosen costs are internally consistent.

Transformation values among character states are, in essence, distances. Distances between sequences are these character changes summed over the length of the sequences. When analyzed as such, they must conform to certain logical strictures even if the events they describe do not. Foremost among these is the triangle inequality invoked by Farris (1981, 1985) and others (Swofford 1981) in their criticisms of distances. The triangle inequality is a property of metric spaces. Distances that do not conform to this relation are nonmetric and, hence, are internally inconsistent because

1. Key words: triangle inquality, systematics, character analysis, transition-transversion ratio, gap costs.

Address for correspondence and reprints: Ward C. Wheeler, Department of Invertebrates, American Museum of Natural History, Central Park West at 79th Street, New York, New York 10024-5192.

Mol. Biol. Evol. 10(3):707–712. 1993.

0737-4038/93/1003-0016$02.00

they require unobserved transformations. The requirement that this inequality apply to matrix characters is reasonable, because these values are distances.

The Triangle Inequality

The triangle inequality is defined as $d_{ij} \leq d_{ik} + d_{kj}$ for all triplets of character states (i, j, and k). This relationship specifies that the direct path between two character states (e.g., d_{ij}) cannot be longer (more costly) than a less-direct path involving other intermediate states ($d_{ik} + d_{kj}$; fig. 1a, *left*). Without this limitation, it would be not only possible but necessary to postulate unobserved intermediate states between observed terminal states [although this can be forbidden operationally (Williams and Fitch 1989)].

If the costs AC = 3, AG = 1, and GC = 1 are applied, and if only states A and C were observed at a given position, then it would be more parsimonious (i.e., less costly) to postulate an ancestral state (G) not found in any taxon than to reconstruct a path based on observation. I contend that it is illogical to assign unobserved states to hypothetical ancestors.

This approach to character analysis extends the criticisms of distance analysis. Farris (1981, 1985) has pointed out that the violation of this inequality reveals internal inconsistencies in the data. Such violations may be uninterpretable and cannot be understood as evolutionary. In this character-based situation, violation also confers internal inconsistency on the analysis, since the same unreasonable behavior is implied.

The Problem

The analysis of molecular sequence data presents three situations in which values for character transformations, explicitly or implicitly invoked, may violate the triangle inequality: (1) the specification of extreme transition:transversion ratios; (2) the treatment of alignment gaps as missing data; and (3) the use of asymmetrical character-transformation models.

Transition-Transversion Ratio

If a matrix is defined in which only transitions and transversions are distinguished (fig. 1b, *left*), then there is a lower bound on the cost ratio of transversions to transitions. Since a transversion can in some sense subsume a transition (two transversions—e.g., A→C→G—can yield the same result as a single transition—e.g., A→G; fig. 1a, *right*), the cost of a transition must be no more than twice that of a transversion. Otherwise, a transformation between the purines A and G could proceed via the less direct, but less costly, pyrimidine C. Since it is impossible for transversions to proceed via this path (no number of transitions can yield a transversion), there is no theoretical upper limit on the cost of transversions with respect to transitions (although I doubt that the lower limit has been broached in practice).

Treatment of Sequence Gaps

A similar matrix can be created with five states for each nucleotide character (A, C, G, T/U, and a dash (–), which indicates a gap; fig. 1b, *right*). If the transformation cost between a gap and any other character is set too low, then it will be less costly to postulate intermediate gaps between nucleotide transformations. By this logic, hypothetical ancestors would contain only gaps as character states, and any scheme of relationships would have a cost of zero (since all transformations would be from or to gaps). This may seem obvious, but gaps are treated as missing data in most sequence-based analyses (e.g., see Wheeler 1989; Zimmer et al. 1989; Allard and Honeycutt 1992).

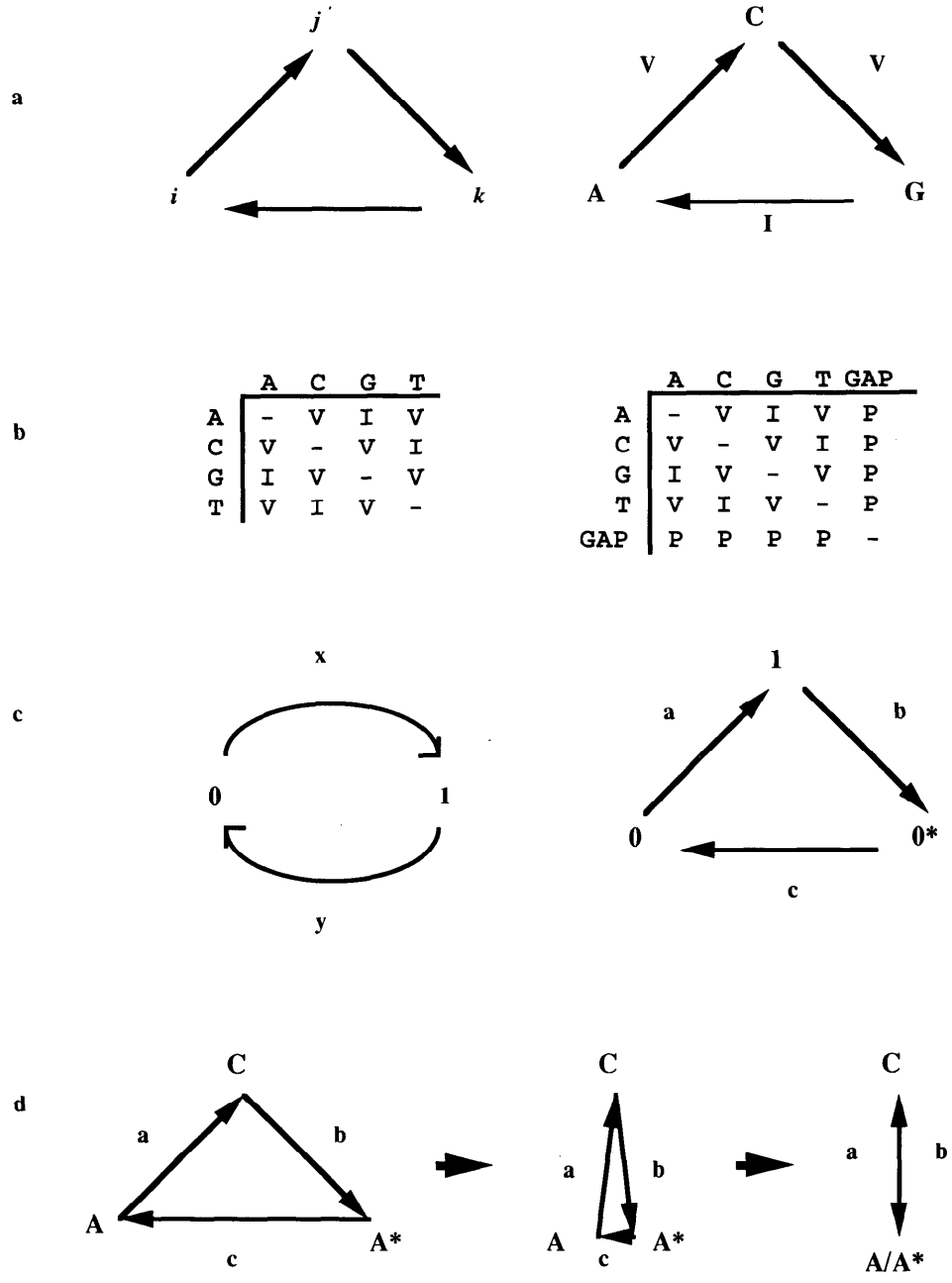

FIG. 1.—Illustrations of the triangle inequality. a (*left*), Geometrical representation of character states i, j, and k. a (*right*), Transformation path involving two transversions, which results in a net transition. b (*left*), Matrix of character-transformation costs for DNA nucleotides. Transitions have cost "I," and transversions have cost "V." b (*right*), Matrix of character-transformation costs for DNA nucleotides and gaps. Transitions have cost "I"; transversions, "V"; and gaps, "P." c (*left*), A binary character (states 0 and 1, as with restriction sites) that shows asymmetrical transformation costs ($x \neq y$). c (*right*), Representation of a binary character with more precise homology statements ($0 \neq 0^*$) and more complete costs (a, b, and c). d, The geometrical limit of a triangle as it progresses to a line ($c \rightarrow 0$). This limit forces the character transformations to be symmetrical, since the length ratio of the nonzero legs (a and b) converges to unity.

Symmetry

Asymmetrical character-transformation costs have been suggested for many situations (Doyle et al. 1990; Swofford and Olsen 1990; Blouin et al. 1992). Foremost among these is the analysis of restriction-site data (fig. 1c, *left*). Because of the means of sampling sequence change, it may be very many times more likely to see the loss of a site than it would be to see its gain (DeBry and Slade 1985) even if the underlying sequence evolution is symmetrical. In brief, a model of evolution with two states is proposed: The two states (1 = presence; and 0 = absence) are linked by two paths, one for each direction of change (x and y). The process of gain and loss would start with state 0, proceed to state 1 with one cost (x), and then back to 0 with another (y). The assumption of this scenario is that the two 0's (primitive and derived) are the same. They are not. While the states may be indistinguishable, this does not mean that they are the same. The 0 states primitive (initial) and derived have unique origins and are potentially differentiable at the sequence level. A more appropriate representation would contain three states 0, 1, and 0*. In this scenario, there are three transformation costs (a, b, and c; fig. 1c, *right*). Cost a corresponds to cost x in the previous arrangement while cost y is split into costs b and c. The argument that $x > y$ can then be reexpressed as $a > b + c$. This is clearly a violation of the triangle inequality, since a must be less than b + c. The use of asymmetrical costs for restriction-enzyme data is therefore incompatible with the basic requirements of the triangle inequality. The cause of its seeming sensibility comes from the initial, mistaken identity $0 = 0^*$.

This argument can be extended to other data as well. Derived nucleotides are not homologous with those present in the ground state, hence the argument is identical (simply substitute A for 0, C for 1, and A* for 0*). Unfortunately, A* can never be observationally distinguished from A (although they can be on a cladogram). This situation can be thought of as the limiting case where c goes to zero. To maintain the triangle relationship, b must converge to a as the triangle progresses to a line (fig. 1d).

$$\lim_{c \to 0} a/b = 1 .$$

This necessary symmetry has several implications. First among these is that asymmetrical character-transformation values do not measure polarity but incorrect putative homology statements. Methods that derive matrices for nucleotide data (e.g., see Wheeler 1990*a*) having asymmetrical costs cannot be used to root cladograms, as I have suggested elsewhere (Wheeler 1990*b*). They can be used, however, to test for unobserved intermediate states in nucleotide change. The higher the asymmetry, the greater the number of unobserved intermediates. This also suggests that the more taxa that are used in a study, the more unlikely it is to see large asymmetries, since the mistaken putative homologies will be discovered through additional observation (homoplasy). The more poorly the group is sampled, the more asymmetrical the data are likely to appear. This requirement of symmetry (a general property of a metric but shown as an expression of the triangle inequality) does not exclude polarity statements such as those derived from ontogeny or outgroups; it merely excludes different costs based on the direction of change.

Conclusion

Character-transformation models, as described by matrices, must be logical. The general properties of a metric proceed directly from the triangle inequality, and only through its application can logical and consistent cost scenarios be created. The three transformation parameter limits proposed here—on transition-transversion ratio, on gap costs, and on symmetry—are often violated but are nonetheless necessary. With

these limits in mind, the next step is to more firmly pare down the possible world of values, in order to yield more appropriate schemes for phylogenetic analysis.

Acknowledgments

I would like to thank Ranhy Bang, James Carpenter, Rob DeSalle, Steven Farris, Walter Fitch, John Gatesy, Cheryl Hayashi, Micheal Novacek, Norman Platnick, Charles Ray, Alfreid Vogler, Paul Vrana, and Michael Whiting for critical and helpful comments during the preparation of the manuscript of this letter.

LITERATURE CITED

ALLARD, M. W., and R. L. HONEYCUTT. 1992. Nucleotide sequence variation in the mitochondrial 12S rRNA gene and the phylogeny of African mole-rats (Rodentia: Bathyergidae). Mol. Biol. Evol. **9**:27–40.

BLOUIN, M. S., J. B. DAME, C. A. TARRANT, and C. H. COURTNEY. 1992. Unusual population genetics of a parasitic nematode: mtDNA variation within and among populations. Evolution **46**:470–476.

BROWN, W. M., E. M. PRAGER, A. WANG, and A. C. WILSON. 1982. Mitochondrial DNA sequences of primates: tempo and mode of speciation. J. Mol. Evol. **18**:225–239.

CAMIN, J. H., and R. R. SOKAL. 1965. A method for deducing branching sequences in phylogeny. Evolution **19**:311–326.

DEBRY, R. W., and N. A. SLADE. 1985. Cladistic analysis of restriction endonuclease cleavage maps within a maximum-likelihood framework. Syst. Zool. **34**:21–34.

DOYLE, J. J., J. L. DOYLE, and A. H. D. BROWN. 1990. A chloroplast-DNA phylogeny of the wild perennial relatives of soybean (Glycine subgenus Glycine): congruence with morphological and crossing groups. Evolution **44**:371–389.

FARRIS, J. S. 1977. Phylogenetic analysis under Dollo's law. Syst. Zool. **26**:77–88.

———. 1981. Distance data in phylogenetic analysis. Pp. 3–22 *in* V. A. FUNK and D. R. BROOKS, eds. Advances in cladistics: proceedings of the first meeting of the Willi Hennig Society. New York Botanical Garden, New York.

———. 1985. Distance data revisited. Cladistics **1**:67–85.

———. 1988. HENNIG86, version 1.5. Port Jefferson, N.Y.

FITCH, W. M. 1971. Toward defining the course of evolution: minimum change for a specific tree topology. Syst. Zool. **20**:406–416.

LIU, H., and A. T. BECKENBACH. 1992. Evolution of the mitochondrial cytochrome oxidase II gene among 10 orders of insects. Mol. Phylogenet. Evol. **1**:41–52.

SANKOFF, D. D., and R. J. CEDERGREN. 1983. Simultaneous comparison of three or more sequences related by a tree. Pp. 253–264 *in* D. D. SANKOFF and J. B. KRUSKAL, eds. Time warps, string edits, and macromolecules: the theory and practice of sequence comparison. Addison-Wesley, Reading, Mass.

SANKOFF, D. D., and P. ROUSSEAU. 1975. Locating the vertices of a Steiner tree in arbitrary metric space. Math. Prog. **9**:240–246.

SWOFFORD, D. L. 1981. On the utility of the distance Wagner procedure. Pp. 25–43 *in* V. A. FUNK and D. R. BROOKS, eds. Advances in cladistics: proceedings of the first meeting of the Willi Hennig Society. New York Botanical Garden, New York.

———. 1990. PAUP: phylogenetic analysis using parsimony, version 3.0q. Illinois Natural History Survey, Champaign.

SWOFFORD, D. L., and G. J. OLSEN. 1990. Phylogeny reconstruction. Pp. 411–502 *in* D. M. HILLIS and C. MORITZ, eds. Molecular systematics. Sinauer, Sunderland, Mass.

WHEELER, W. C. 1989. The systematics of insect ribosomal DNA. Pp. 307–321 *in* B. FERNHOLM, K. BREMER, L. BRUNDIN, H. JÖRNVALL, L. RUTBERG, and H.-E. WANNTORP, eds. The hierarchy of life. Elsevier, Amsterdam.

———. 1990*a*. Combinatorial weights in phylogenetic analysis: a statistical parsimony procedure. Cladistics **6**:269–278.

———. 1990*b*. When is an outgroup not an outgroup and how to root DNA sequence based topologies without an outgroup. Cladistics **6**:363–367.

WILLIAMS, P. L., and W. M. FITCH. 1989. Finding the minimal change in a given tree. Pp.

453–470 *in* B. FERNHOLM, K. BREMER, L. BRUNDIN, H. JÖRNVALL, L. RUTBERG, and H.-E. WANNTORP, eds. The hierarchy of life. Elsevier, Amsterdam.

ZIMMER, E. A., R. K. HAMBY, M. L. ARNOLD, D. A. LEBLANC, and E. C. THERIOT. 1989. Ribosomal RNA phylogenies and flowering plant evolution. Pp. 205–214 *in* B. FERNHOLM, K. BREMER, L. BRUNDIN, H. JÖRNVALL, L. RUTBERG, and H.-E. WANNTORP, eds. The hierarchy of life. Elsevier, Amsterdam.

WALTER M. FITCH, reviewing editor

Received October 9, 1992; revision received January 15, 1993

Accepted January 15, 1993