DEALING WITH INCOMPLETENESS: NEW ADVANCES FOR THE USE OF FOSSILS IN PHYLOGENETIC ANALYSIS

Author(s): IGNACIO H. ESCAPA and DIEGO POL
Source: Palaios, 26(3):121-124. 2011.
Published By: Society for Sedimentary Geology
URL: http://www.bioone.org/doi/full/10.2110/palo.2011.S02

PALAIOS
Emphasizing the impact of life on Earth's history

PALAIOS, 2011, v. 26, p. 121–124
Spotlight
DOI: 10.2110/palo.2011.S02

SPOTLIGHT

DEALING WITH INCOMPLETENESS: NEW ADVANCES FOR THE USE OF FOSSILS IN PHYLOGENETIC ANALYSIS

IGNACIO H. ESCAPA* and DIEGO POL
CONICET-Museo Paleontológico Egidio Feruglio, Trelew, Chubut, 9100, Argentina
e-mail: iescapa@mef.org.ar

The importance of fossils in understanding the evolutionary history of organisms was a controversial topic of debate during the first few decades in the history of phylogenetic systematics. During this time some authors suggested that extinct taxa could have only a minor role in phylogeny reconstruction (e.g., Patterson, 1981). For the most part, these types of bold statements were based on the fact that fossils are usually incomplete and, therefore, presumably not capable of overturning hypotheses based on the wealth of phylogenetic information that extant taxa provide. However, phylogenetic studies based only on extant organisms are certainly missing a large part of the diversity that arose during the evolutionary history of a taxonomic group and therefore use a highly biased sampling of the available information.

The short history of phylogenetics, however, has shown that many heated discussions are put to rest with empirical data rather than by rhetorical debates. After several empirical studies clearly showed the critical role of fossils for understanding the evolution of major groups of organisms (Gauthier et al., 1988; Donoghue et al., 1989), advocates of ignoring fossil taxa in phylogenetic reconstruction rapidly disappeared (at least from the literature). The major reason for the importance of fossils in phylogenetic reconstruction is that they can bear unique combinations of characters, which are absent in extant taxa, and these can be critical to test the interrelationships of all analyzed species. This, in fact, is not a property exclusive of extinct organisms and the inclusion of extant taxa can be equally critical and necessary, depending on the characteristics of the phylogenetic problem being analyzed. In general terms, the importance of extinct taxa in phylogenetic reconstruction stems from the importance of achieving a taxon sampling scheme as complete as possible for solving a phylogenetic problem. In recent years increasing the taxon sampling has been shown to improve the performance of phylogenetic methods (Zwickl and Hillis, 2002; Heath et al., 2008).

If we consider that virtually all the species of plants and animals that ever lived are now extinct (Raup, 1986), ignoring fossil biodiversity will clearly lead to a remarkably biased design of the taxonomic sampling scheme used in a phylogenetic analysis. Moreover, such design will determine the inclusion or exclusion of a given lineage based on contingencies in the history of a lineage (e.g., becoming extinct) rather than based on whether or not it bears a unique combination of characters that may be relevant for solving a phylogenetic problem. In sum, as Cobbett et al. (2007) recently demonstrated, fossils are not a special class of organisms but may provide pieces of information that are as relevant as any other taxon; therefore, their exclusion from the data analysis seems completely unjustified.

The inclusion of fossils, however, does bring into a phylogenetic analysis a special bonus, their geologic age. The first appearance datum of a given species or lineage provides a direct way to infer the (minimum) age of origin of a given clade. This is a type of so-called tree-based inference and is obviously contingent on: (1) an accurate determination of the age of the fossil-bearing deposits and (2) the phylogenetic position of the fossils in the cladograms (Pol and Norell, 2006). The inclusion of fossils in phylogenetic analyses, therefore, provides a rich source of information to understand the tempo and mode of the evolutionary history of a group, dating major biotic events (Crepet et al., 2004; Hermsen and Hendricks, 2006) both from a direct point of view and through calibrations of molecular clocks (Sanderson, 2003).

Most authors accept that both extinct and extant taxa are desirable for inclusion to understand the phylogenetic history of a group. We think, however, that there are several unfounded fears widespread in the phylogenetic community, most of which are related to the quality and quantity of phylogenetically adequate information that can be retrieved from extinct organisms. Here we review some recent

Ignacio Escapa (right) received his Ph.D. from the Universidad Nacional del Comahue (Bariloche, Argentina) under Rubén Cúneo, studying the Jurassic Floras of Central Patagonia (Cañadón Asfalto and Cañadón Calcareo Formations). After his Ph.D., he joined Thomas N. Taylor and Edith L. Taylor in the Department of Ecology and Evolutionary Biology (University of Kansas) for a one-year postdoctoral fellowship studying particular aspects of the Triassic floras of Antarctica. His research interests include Jurassic and Triassic floras of Gondwana, conifer phylogeny, and methods for dealing with fossil plants in phylogenetic studies. Currently he is a CONICET (Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina) postdoc at the Museo Paleontológico Egidio Feruglio (Trelew, Argentina) and works on the diversity and evolution of the Jurassic floras of Patagonia.

Diego Pol (left) received his M.S. and Ph.D. at the joint program of the American Museum of Natural History and Columbia University (New York) with Mark Norell, studying the phylogenetic relationships of basal sauropodomorph dinosaurs and methods for combining stratigraphic and phylogenetic information. Currently he is a CONICET researcher at the Museo Paleontológico Egidio Feruglio and works on several projects related to understanding the evolution of archosaur reptiles from the Mesozoic of South America, with special focus on the highly diverse Jurassic fauna of central Patagonia.

* Corresponding author.

 0883-1351/11/0026-0121/$3.00

methodological advances and developments that are related to these problems and help in retrieving useful information from fossils in phylogenetic analyses.

MISSING ENTRIES: TAXON INSTABILITY, CONSENSUS TREES, AND COMPUTATIONAL ADVANCES

In comparison with extant taxa, fossils usually have much less information that can be scored for phylogenetic characters, which results in a larger number of missing entries in the data matrices. Several authors have noted the negative effects of including taxa with abundant missing entries—taxa known from poorly preserved or incomplete specimens—that mostly boil down to obtaining an exceedingly large number of most parsimonious trees, given that fragmentary taxa may take multiple and equally optimal alternative positions (Gauthier, 1986; Novacek, 1992; Wilkinson and Benton, 1995). This usually leads to a poorly resolved strict consensus tree. Numerous authors, therefore, have applied a simple criterion for excluding fossil taxa from a phylogenetic analysis based on establishing a cut-off level for percentage of missing entries, i.e., if a taxon has >75% of missing entries it should be excluded from the data matrix, either implicitly or explicitly (e.g., Grande and Bemis, 1998). Recent advances in phylogenetic software currently allow for the analysis and comparison of thousands of equally parsimonious trees in an efficient manner (e.g., TNT; Goloboff et al., 2008). Wilkinson (1994) and Wilkinson et al. (2004) have pioneered the development of consensus methods that can extract phylogenetic information common to all MPT (most parsimonious trees) and summarize this information in what they term reduced consensus trees—highly resolved consensus trees that ignore the alternative positions of unstable taxa. Kearney (2002) rightfully pointed out that instability can be caused by a high number of missing entries but can also be due to the presence of character conflict. A recent protocol, IterPCR (Pol and Escapa, 2009), combines these two approaches and allows the automatic identification of unstable taxa as well as the characters responsible for taxon instability within the context of a cladistic analysis of a group (Fig. 1). This method not only produces a highly resolved reduced consensus but also provides for the user a list of characters that may be related to the instability of a given taxon. This method differentiates the set of characters that positively support alternative positions of the unstable taxon in the MPTs (i.e., those showing character conflict) from the set of characters scored with missing entries that may diminish the instability of each taxon if the condition were known.

TESTING HOMOLOGY HYPOTHESES IN A PARSIMONY CONTEXT: DYNAMIC HOMOLOGY APPROACH

The definition of morphological characters is one of the most basic theoretical and empirical issues of phylogenetic systematics. Characters are organized in a conceptual grid of homology correspondences, classically defined as primary homologies (de Pinna, 1991). The assessment of the various primary homology hypotheses is based on such sources of information as comparative anatomy and development, and is also influenced by previous knowledge on the relationships of the group being analyzed.

In some cases, more than one alternative hypothesis of primary homology can be postulated for a particular structure. Having multiple hypotheses of primary homology poses a methodological challenge to the traditional approach of phylogenetic analysis based on morphological data. This is relatively frequent in comparative studies of fossil taxa given that preservational causes restrict the amount of information that can be gathered (Nixon, 1996). Given the lack of a proper phylogenetic method to evaluate competing hypotheses of primary homology, discussions of alternative correspondences among particular structures were often tautological. Classic examples of this are the homology relationships of the angiosperm flower (e.g., Doyle, 2006) and the manual digits in the dinosaur-bird transition (e.g., Xu et al., 2009). Ramirez (2007) proposed a quantitative method to evaluate alternative homology hypotheses within a parsimony context, which is analogous to the dynamic homology procedure used for aligning DNA sequences (Wheeler et al., 2006). In this context, the alternative hypotheses of homology for a given structure are tested by their congruence with the rest of the characters included in the matrix. Using the dynamic homology approach, the parsimony analysis is applied not just to find the MPT, but also to determine the most parsimonious homology hypothesis for the structure under study. The concept of dynamic homology, when applied to morphological characters, constitutes a recent approach, and several methodological and theoretical issues still need to be analyzed (Agolin and D'Haese, 2009); however, this clearly represents the most realistic method so far proposed for testing hypotheses of primary homology.

MORE INFORMATION FROM THE SAME CHARACTERS: MORPHOMETRICS AND CLADISTICS

Since the early development of phylogenetic systematics, most analyses that included only morphology used discrete and meristic (quantitatively defined) characters, probably because this type of data was the only one supported by most phylogenetic software. Systematists working at low taxonomic levels, however, usually find morphological differences among the studied organisms that may be better analyzed quantitatively than qualitatively (Fig. 2). Thus, integrative phylogenetic analyses aiming to simultaneously solve problems at different taxonomic levels will require the combined use of discrete and continuous characters in order to include all the relevant information. Molecular-based phylogenetic studies show a similar situation, considering the different rates of evolution of different genes (Giribet, 2002, 2010).

The development of efficient methods to recover phylogenetic information from quantitatively varying structures is particularly relevant when including fossil taxa in the analysis, since available data are limited in different degrees. Goloboff et al. (2006) implemented algorithms in the software TNT (Goloboff et al., 2008) for the analysis of continuous characters as additive characters, allowing the exploration of several new approaches in cladistic analysis. This development has allowed several empiric studies to explore the use of continuous characters without confining the variation into discrete units (Pereyra and Mound, 2009; Vega et al., 2009), which usually relied on arbitrary or loosely justified criteria. The direct implementation of continuous characters is useful to consider variation in such unidimensional space as direct measurements or ratios. To be analyzed in a quantitative way, however, shape traits require the use of tools that permit analysis in a multivariate space. Gonzalez-José et al. (2008) proposed a combination of geometric morphometrics (GM) and cladistics in order to develop the phylogeny of genus *Homo*. In this approach, GM is used to capture the shape change in particular characters and the resultant principal components (PC) are used as continuous characters in the construction of a cladistic matrix. By using this protocol, more phylogenetic information from the same structures can be recovered and hypothetical ancestral character states for each of the lineages can be reconstructed. Even when this approach is powerful in terms of recovering phylogenetic information enclosed in multidimensional shapes, the use of PC shows a number of biases that can potentially affect the results. In this context, Catalano et al. (2010) have developed a method for the use of landmark position data in phylogenetic analysis, wherein changes in the relative position of the individual landmarks represent shape changes in a particular structure.

REMARKS AND FUTURE DIRECTIONS

Several problems associated with the limited phylogenetic information that can be retrieved from fossil taxa are being overcome with

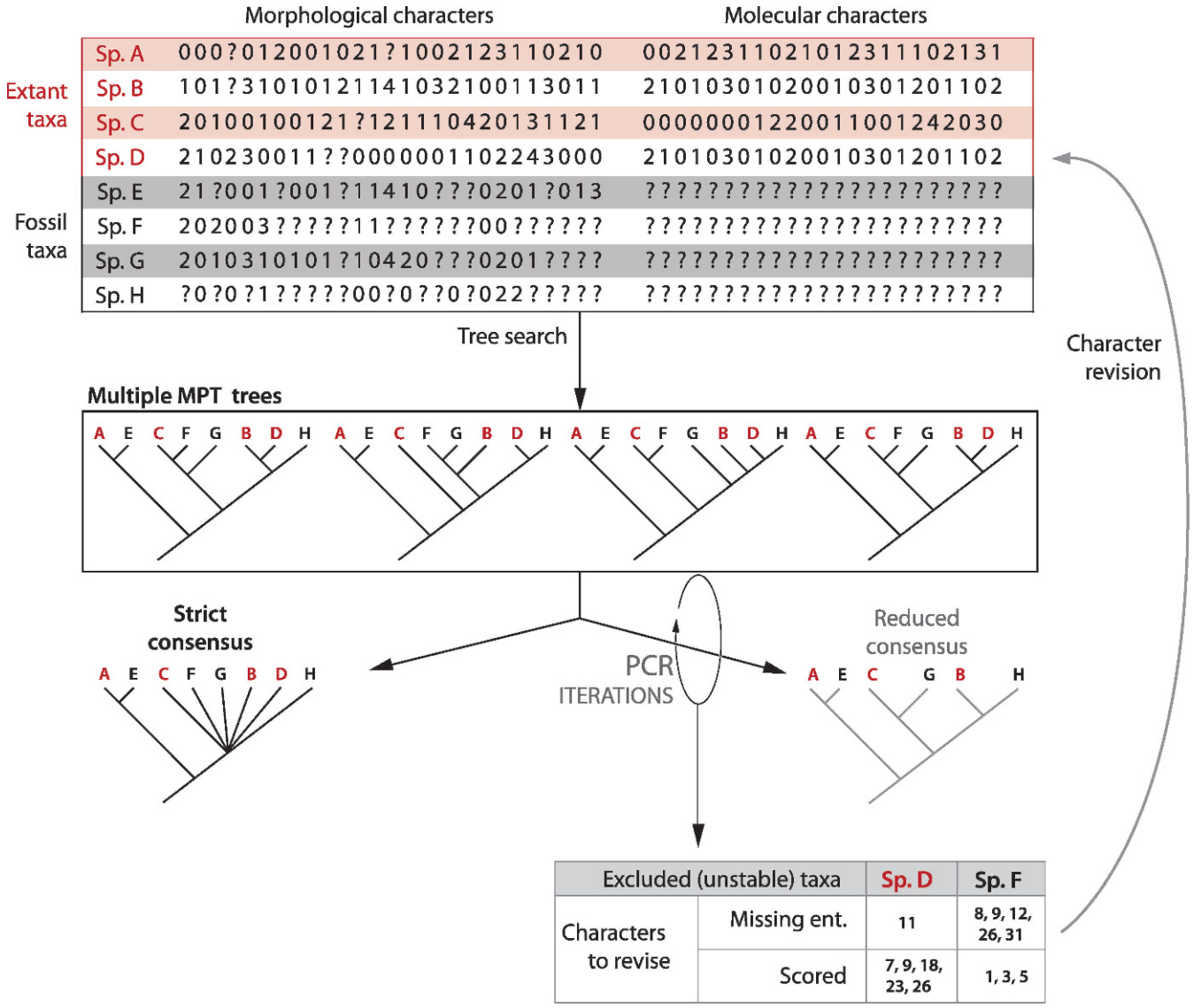

FIGURE 1—Workflow of data analysis and revision using the IterPCR procedure (Pol and Escapa, 2009). The data matrix includes both extant and extinct taxa and morphological and molecular characters. The tree search produces a collection of most parsimonious trees that can be summarized in different ways. The strict consensus tree can be poorly resolved when unstable taxa are present. Through the IterPCR procedure, a highly resolved, reduced consensus tree and a chart of the characters related to the taxon instability is produced.

methodological developments and improvements in computational implementations. This means that researchers can now efficiently analyze fossils of uncertain phylogenetic position without eliminating them from the analysis (IterPCR) or effectively test alternative hypotheses of homology within a phylogenetic framework for fossil structures that are difficult to interpret due to preservation biases or evolutionary transformations. The incorporation of geometric morphometric data into phylogenetic analysis opens a wide range of possibilities to analyze shape changes from an evolutionary perspective. All of these advances have been developed with a common objective—maximizing the use of phylogenetic data by avoiding the *a priori* elimination of data, observations, or hypotheses in phylogenetic analysis. Having the tools to deal with problematic entities (characters or taxa) makes testing the influence of information from extinct organisms on phylogenetic relationships easier.

The issues highlighted here represent areas of active research and there are certainly numerous problems that still need to be solved. For instance, unstable or highly incomplete fossil taxa usually lead to low values of measures of nodal support which obscure the robustness of a phylogenetic analysis (Wilkinson et al., 2000). Methods and efficient computational implementations that explore this issue will certainly help enormously to obtain a deeper understanding of the phylogenetic information of a given dataset, with a thorough assessment of its weaknesses and robustness.

In the case of morphometrics and cladistics, further research will be directed to combining the use of this kind of data with other sources of information; e.g., classic morphological characters, and molecular data. Simultaneous analysis combining these sources of data creates such methodological problems as the differential weighting of characters. In addition, the concept of multivariate characters requires the development of new tools to estimate character independence. Future developments in the field of morphological dynamic homology will likely bring increased capabilities of this method for the simultaneous analysis of conflicting homology statements using multiple structures at the same time. This will require an efficient design in the computational implementation of this approach as the complexity of the problem can increase combinatorially with the number of conflicting homology hypotheses and the number of structures.

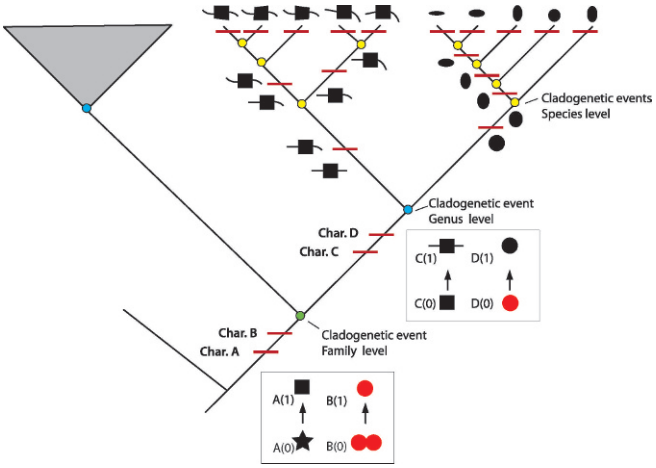

FIGURE 2—Hypothetical phylogenetic tree showing the evolution of two features through different taxonomic levels. Changes occurring at basal nodes can be considered as discrete (e.g., absence or presence of particular structures, major changes of shape and color, etc). At the terminal nodes (species level cladogenetic events), however, the changes undergone by the structures under consideration are better explained in a quantitative way. Boxes show transformations in two hypothetical characters.

REFERENCES

Agolin, M., and D'Haese, C., 2009, An application of dynamic homology to morphological characters: Direct optimization of setae sequences and phylogeny of the family Odontellidae (Poduromorpha, Collembola): Cladistics, v. 25, p. 353–385.

Catalano, S., Goloboff, P., and Giannini, N., 2010, Phylogenetic morphometrics (I): The use of landmark data in a phylogenetic framework: Cladistics, v. 26, p. 1–11.

Cobbett, A., Wilkinson, M., and Wills, M.A., 2007, Fossils impact as hard as living taxa in parsimony analyses of morphology: Systematic Biology, v. 56, p. 753–766.

Crepet, W.L., Nixon, K.C., and Gandolfo, M.A., 2004, Fossil evidence and phylogeny: The age of major angiosperm clades based on mesofossil and macrofossil evidence from Cretaceous deposits: American Journal of Botany, v. 91, p. 1666–1682.

De Pinna, M.C.C., 1991, Concepts and tests of homology in the cladistic paradigm: Cladistics, v. 7, p. 367–394.

Donoghue, M.J., Doyle, J.A., Gauthier, J., Kluge, A.G., and Rowe, T., 1989, The importance of fossils in phylogeny reconstruction: Annual Review of Ecology and Systematics, v. 20, p. 431–460.

Doyle, J.A., 2006, Seed ferns and the origin of angiosperms: Journal of the Torrey Botanical Society, v. 133, p. 169–209.

Gauthier, J.A., 1986, Saurischian monophyly and the origin of birds: Memoirs of the California Academy of Sciences, v. 8, p. 1–47.

Gauthier, J.A., Kluge, A.G., and Rowe, T., 1988, Amniote phylogeny and the importance of fossils: Cladistics, v. 4, p. 105–209.

Giribet, G., 2002, Current advances in the phylogenetic reconstruction of metazoan evolution. A new paradigm for the Cambrian explosion?: Molecular Phylogenetics and Evolution, v. 24, p. 345–357.

Giribet, G., 2010, A new dimension in combining data? The use of morphology and phylogenetic data in metazoan systematics: Acta Zoologica, v. 91, p. 11–19.

Goloboff, P.A., Farris, J.S., and Nixon, K.C., 2008, TNT, a free program for phylogenetic analysis: Cladistics, v. 24, p. 774–786.

Goloboff, P.A., Mattoni, C.I., and Quinteros, S., 2006, Continuous characters analyzed as such: Cladistics, v. 22, p. 1–13.

González-José, R., Escapa, I., Neves, W., Cúneo, R., and Pucciarelli, N., 2008, Cladistic analysis of continuous modularized traits provides phylogenetic signals in *Homo* evolution: Nature, v. 453, p. 775–778.

Grande, L., and Bemis, W.E., 1998, A comprehensive phylogenetic study of amiid fishes (Amiidae) based on comparative skeletal anatomy. An empirical search for interconnected patterns of natural history: Journal of Vertebrate Paleontology, v. 18, p. 1–690.

Heath, T.A., Zwickl, D.J., Kim, J., and Hillis, D.M., 2008, Taxon sampling affects inferences of macroevolutionary processes from phylogenetic trees: Systematic Biology, v. 57, p. 160–166.

Hermsen, E.J., and Hendricks, J.R., 2006, The hierarchy of time: PALAIOS, v. 21, p. 403–405.

Kearney, M., 2002, Fragmentary taxa, missing data, and ambiguity: Mistaken assumptions and conclusions: Systematic Biology, v. 51, p. 369–381.

Nixon, K.C., 1996, Paleobotany in cladistics and cladistics in paleobotany: Enlightenment and uncertainty: Review of Palaeobotany and Palynology, v. 90, p. 361–373.

Novacek, M.J., 1992, Fossils, topologies, missing data, and the higher level phylogeny of eutherian mammals: Systematic Biology, v. 41, p. 58–73.

Patterson, C., 1981, Significance of fossils in determining evolutionary relationships: Annual Review of Ecology and Systematics, v. 12, p. 195–223.

Pereyra, V., and Mound, L., 2009, Phylogenetic relationships within the genus *Cranothrips* (Thysanoptera, Melanthripidae) with consideration of host associations and disjunct distributions within the family: Systematic Entomology, v. 34, p. 151–161.

Pol, D., and Escapa, I.H., 2009, Unstable taxa in cladistic analysis: Identification and the assessment of relevant characters: Cladistics, v. 25, p. 515–527.

Pol, D., and Norell, M., 2006, Uncertainty in the age of fossils and the stratigraphic fit for phylogenies: Systematic Biology, v. 55, p. 512–521.

Ramirez, M.J., 2007, Homology as a parsimony problem: A dynamic homology approach for morphological data: Cladistics, v. 23, 588–612.

Raup, D.M., 1986, Biological extinction in earth history: Science, v. 231, p. 1528–1533.

Sanderson, M.J., 2003, R8s: Inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock: Bioinformatics, v. 19, p. 301–302.

Vega, A., Rua, G.H., Fabbri, L.T., and Rúgolo de Agrásar, Z.E., 2009, A morphology-based cladistic analysis of *Digitaria* (Poaceae, Panicoideae, Paniceae): Systematic Botany, v. 34, p. 312–323.

Wheeler, W.C., Aagesen, L., Arango, C.P., Faivovich, J., Grant, T., D'Haese, C., Janies, D., Smith, W.L., Varon, V., and Giribet, G., 2006, Dynamic Homology and Phylogenetic Systematics: A Unified Approach Using POY: American Museum of Natural History Press, New York, 357 p.

Wilkinson, M., 1994, Common cladistic information and its consensus representation: Reduced Adams and reduced cladistic consensus trees and profiles: Systematic Biology, v. 43, p. 343–368.

Wilkinson, M., and Benton, M.J., 1995, Missing data and rhynchosaur phylogeny: Historical Biology, v. 10, p. 137–150.

Wilkinson, M., Cotton, J., and Thorley, J., 2004, The information content of trees and their matrix representations: Systematic Biology, v. 53, p. 989–1001.

Wilkinson, M., Thorley, J.L., and Upchurch, P., 2000, A chain is no stronger than its weakest link: Double decay analysis of phylogenetic hypotheses: Systematic Biology, v. 49, p. 754–776.

Xu, X., Clark, J.M., Mo, J., Choiniere, J., Forster, C.A., Erickson, G.M., Hone, D.W.E., Sullivan, C., Eberth, D.A., Nesbitt, S., Zhao, Q., Hernández, R., Jia, C.-K., Han, F.-L., and Guo, Y., 2009, A Jurassic ceratosaur from China helps clarify avian digital homologies: Nature, v. 459, p. 940–944.

Zwickl, D.J., and Hillis, D.M., 2002, Increased taxon sampling greatly reduces phylogenetic error: Systematic Biology, v. 51, p. 588–598.