doi: 10.26492/gbs71(2).2019-04

Information content for biological classifications

T.F. Stuessy

Herbarium and Department of Evolution, Ecology, and
Organismal Biology, The Ohio State University, 1315 Kinnear Road,
Columbus, Ohio 43212, U.S.A.
stuessy.1@osu.edu
Department of Botany and Biodiversity Research,
University of Vienna, Rennweg 14,
A-1030 Vienna, Austria.

ABSTRACT. Classification is a fundamental activity of the human species. The aim of all forms of classification is to establish a hierarchical structure of information that serves as a reference system to answer specific questions. In biological classification the objective is to store data in a conveniently retrievable fashion, to infer evolutionary relationships, and to predict undocumented characteristics of the included organisms. Different kinds of data have been used to form a basic data matrix from which to construct biological classifications. Dendrograms have been traditionally used to illustrate relationships among taxa, although such two-dimensional diagrams do not capture all relationships from the original data matrix. Controversies have existed on which algorithms are best suited to construct dendrograms. Explicit phyletic (evolutionary), phenetic, and cladistic schools of quantitative classification have each offered methods for doing do, and each has made claims for capturing maximum information. Decisions on which type of data and algorithms to use depend upon the nature of the systematic and evolutionary questions being posed. Important is the need for detailed evolutionary investigations so that inferred relationships can be properly evaluated. Information theory, a separate discipline, is viewed as having high potential to enrich information content of biological classifications.

Keywords. Algorithms, characters, data matrix, evolution, information, phylogeny

Introduction

Biological classifications are the cornerstones of ordering and understanding biodiversity. It is uncertain how many species of organisms may inhabit the Earth, but estimates have pointed to at least 8.7 million in total (Mora et al., 2011), with 1.7 million having been already formally described and classified (IUCN, 2010). Our initial challenge is to gain an understanding of the existence of all organisms as a step toward clarifying our world. With this knowledge, we might hope for better management of our biotic resources, perhaps enabling our own future survival. One can hardly be successful with management of a resource if the level of knowledge about it is strikingly incomplete. Biological classifications serve as reference systems for the storage and retrieval of information about these organisms. This system, based on similarities and/or differences, yields a structure of information that places each organism into a specific character space. This allows us to find these organisms within the structure and to deal with them in whatever manner is needed, such as further studies on their reproductive biology, investigating their biogeography, potentials for cultivation, etc.

The placement of organisms in a structure of information represented by classification provides us the ability to predict features not previously used for initial classification or that were never investigated. Classifications with the highest degrees of prediction allow maximum efficiency and precision in the search for new information. A dramatic example of such a potential is illustrated by searches for new medicines from natural plant products, also known as "biological prospecting" (Miller, 1996; Moran et al., 2001). If, for example, a potent alkaloid is extracted from a species of flowering plant and found to be active against some bacterial disease or malignant tumour, it would be most efficient to examine related species in the same genus for other alkaloids that may be equally or more potent against the malady. Without the structure of information provided by classification, we would be reduced to sampling one by one all the c. 350,000 species of flowering plants (Govaerts, 2001, 2003; Bramwell, 2002; Thorne, 2002; Scotland & Wortley, 2003). The cost of this would be so great, not to mention unfeasible within a reasonable period of time, that the task would simply not be done. For the use of the biotic world for human needs, we require classification. Society, in fact, recognises this need, and this is the main reason systematic biology exists as a supported field of human inquiry.

The predictive quality of a classification is dependent upon the degree of information within it. The more accurate and abundant the information that supports the structure of a classification, the greater will be its predictive efficacy. The challenge with constructing classifications with maximum information content is that different types of information exist. Data about organisms can be gathered through description and/or measurement and placed in a basic data matrix to ensure completeness and to facilitate quantitative comparison that allows groups to be formed and ranked hierarchically. Which type of data should be selected (i.e. morphology, cytology, nucleotide sequences, etc.), how these should be divided into characters and states, and how they should be compared, are major challenges.

Different types of information may serve to more appropriately answer different kinds of systematic questions (Stuessy, 2013). For example, data believed useful for asking questions regarding phylogenetic relationships among families of angiosperms may not be suitable for examining the dynamic of interspecific natural hybridisation. Phylogenetic analyses at the infraspecific populational level will likely require data different from those useful for examining biogeographic relationships among genera impacted by the breakup of Gondwanaland.

The purposes of this paper, therefore, are to: (1) review briefly the different types of information that can be used for purposes of constructing biological classifications; (2) discuss approaches for the distillation of information from the basic data matrix; and (3) sketch the synthesis of information for answering different kinds of systematic questions.

Different types of biological information

Classification is based on a comparison of data (information) to form units of organisms. For maximum predictive quality, this involves grouping and ranking in a hierarchical structure. It is possible to construct a classification with only coordinate units, such as is achieved with statistical ordination, but without subordination of groups into subgroups in a hierarchy, the information content of the classification remains extremely low. Ranking of groups, whether informal or formal, is a requirement for predictive classification.

The first step in assembling information for classification is the construction of the basic data matrix. This is the stage of examining the organisms and deciding what types and amounts of information are needed for the study being accomplished. Traditional revisionary systematists may not prepare such a matrix explicitly, or if one is prepared, it may not be published. At minimum, the systematist will select characters and states, intuitively and rapidly, to be used for comparisons among entities (OTUs) for making a classification. Many studies are now based on quantitative assessments of relationships, and therefore, preparation of a basic data matrix is commonplace. How many characters to use and how to relate states have led to much discussion in the literature (see detailed analysis by Soltis, 2014). The more complete the matrix is, the higher will be the level of information available for constructing classifications. Numerous studies have examined the effects of missing data (Maddison, 1993; García-Laencina et al., 2010; Wiens & Morrill, 2011; Brown et al., 2012). Which data and how much can be missing for only minimally disrupting predictive classification depends upon the particular group, algorithm, and questions being asked. The bottom line is: the more complete the data, the better.

The conceptualisation of data into characters and states for the basic data matrix may or may not be challenging depending upon the type of data. It can be relatively straight-forward with nucleotide data, with each base-pair site being a character and the bases being the four states. Gene-order and other molecular data can also be used, which provides more complexity. With morphology or other structural data, the decisions on how to deal with characters and states are very challenging. Studies by Stevens (1991), Hawkins (2000), and Reid & Sidwell (2002) have shown subjectivity in selection of states, even by experienced workers. Nonetheless, for quantitative approaches to classification, these decisions must be taken, and a comprehensive data matrix must be prepared.

Which types of data to include in the matrix to yield the most predictive classification has been debated endlessly through the different data-gathering phases in systematic biology over the past 50 years. Morphology, cytology, secondary plant products, isozymes, and nucleotides have all been championed as the best source of data for general-purpose classification (see citations in Stuessy, 2009a). More recently has been the molecules vs. morphology discussion (Patterson, 1987; Systma, 1990; Patterson et al., 1993; Scotland et al., 2003). Data for the matrix must be selected for specific purposes, i.e. to seek answers appropriate to the questions being asked. Data that seem to have no relevance to phylogeny, e.g. wide-ranging dysploid chromosome numbers that are independent of morphological boundaries, would be unsuitable for making interpretations for reconstructing a phylogenetic diagram. Likewise, to interpret broad-scale genetic trends at the populational level would require population genetic markers (AFLPs, SSRs, RADseq, etc.) and not embryological or anatomical data that tend to be quite conservative and useful mostly at the higher levels of the hierarchy. For interest in constructing a classification for purposes of understanding adaptations to the environment, morphology must be examined. On the other hand, a study emphasising phylogenetic relationships among families of angiosperms within a single order will most probably require nucleotide sequences. Morphology can be helpful here, but rampant parallelism among flowering plants confounds finding correct phylogenetic signal. For example, character states such as inferior ovaries cannot be expected to reveal useful evolutionary information on relationships across all angiosperms because this feature has originated in parallel numerous times (Grant, 1950), and the condition is also somewhat structurally complex (Soltis et al., 2003).

Distillation of information from the basic data matrix

A fundamental approach to distilling evolutionary data from the basic data matrix involves phylogenetic comparisons. The graphic results of such comparisons are often presented in dendrograms, usually rooted or sometimes presented as an unrooted network. The kinds of information that can be inferred in the interpretation of phylogeny are: branching patterns; change of character states within a lineage; number of character states supporting each node; and distinctiveness and cohesiveness of each lineage relative to each other. All of these dimensions are contained in phylogeny reconstruction, but emphasis historically has been placed on the branching patterns, presumably due to the convenience of unambiguously converting such a hierarchical diagram to a hierarchical classification and back again (Stuessy, 2013).

There is no theoretical reason why phylogenetic relationships must be presented graphically in the form of a tree (dendrogram), but ease of understanding affinities and convenience in converting such a diagram into a hierarchical classification have encouraged their use. The tree-making tradition in systematic biology has a long history extending back to Darwin and even earlier in a non-evolutionary context (Voss, 1952; Pietsch, 2012). Construction of the branching diagram is based on some method of inference, which nowadays involves parsimony, maximum likelihood, or Bayesian inference (Baum & Smith, 2012; Stuessy et al., 2014a). Taxa placed close together on the tree are judged to be more closely related than those placed further away. Most cladists have judged the total information content of a tree (or part of a tree) to be the sum of its subgroups (Mickevich & Platnick, 1989).

A similar phylogenetic approach to distilling information content from a basic data matrix has been pattern cladistics. Cladistics developed as a means for determining branching patterns of evolution, i.e. one aspect of phylogeny. This objective emphasised ancestral vs. derived morphological character states at its inception (Hennig, 1950, 1966), which were predetermined in development of the data matrix through arguments regarding polarity (Crisci & Stuessy, 1980; Stevens, 1980). Few characters were selected for their presumed efficacy to reveal evolutionary directionality, and comparison among the states led to production of a branching diagram (cladogram). Practitioners of pattern cladistics (e.g. Brady, 1985; Kemp, 1985; Platnick, 1985) chose to interpret the branching diagram as simply a pattern of information rather than a pattern of evolution. In a sense they were completely correct, as interpretations of a branching pattern is only one dimension of phylogeny and hence inappropriate as a complete portrayal of evolution. Despite the rigor of this interpretation, few advocates remain because it seems odd to be selecting characters and states for phylogenetic purposes to then later interpret the branching diagram solely in a non-evolutionary context.

In addition to branching patterns, phylogenetic diagrams can also reveal the number of character states that support each node of the tree. With the case of morphology these can be very few states, leading to the criticism of weak support or even single-character taxonomy, which has long been rejected as an information basis for classification (Davis & Heywood, 1963; Stuessy, 1990). With nucleotide data, however, the support can be strong. For phylogenetic reconstruction, workers often seek nucleotide data over morphology, especially at higher levels of the hierarchy where evolution and interpretations of morphology become increasingly difficult. Statistical measures that assess the robustness (i.e. veracity) of the nodal structure of a diagram do test the stability of nodal support based on the characters and states used. One must be careful, however, because a support measure, such as the bootstrap (Felsenstein, 1985), can show high support for a node that may, in fact, be based on data inappropriate for the organisms or questions involved.

Another measure of information within a phylogeny is the change of character states within lineages, or the patristic distance (Stuessy, 1987, 1997; Stuessy & König, 2008). Such divergence can yield single taxa and lineages that are dramatically different from the parental stock (ancestor). This is often the case with adaptively radiated island taxa that have diverged morphologically from ancestors in continental areas (Stuessy et al., 2014b). This information is frequently neglected in cladistic classification, but it is taken into account in quantitative evolutionary classification (Stuessy, 2009b).

A further type of information contained in the phylogeny is the cohesiveness and distinctiveness of each taxon and lineage from each other (Stuessy, 2013). This category of information was the basis of the data utilised in phenetic analyses (e.g. Sneath & Sokal, 1973) to interpret relationships, but this was done in the context of overall similarity independent of phylogeny, and hence it has not endured as a general purpose approach to biological classification. This type of information, based on selected characters and states of evolutionary import, is a part of the real phylogeny and should be taken into account for aspects of information distillation.

When quantitative approaches to biological classification began with phenetics in the late 1950s and early 1960s (e.g. Sokal & Sneath, 1963), the different algorithms that were being used to synthesise relationships from the information in the basic data matrix often resulted in different hierarchical dendrograms and resultant classifications. This led to mathematical perspectives on measuring retention (or loss) of this information (Rohlf, 1974). One of the commonly used measures of evaluating information transfer was the cophenetic correlation coefficient (Sokal & Rohlf, 1962; Farris, 1969). As cladistics developed in the 1970s and 1980s, many studies have attempted to measure the information content of cladograms in comparison with the original data matrix. Most commonly used have been the consistency index (Kluge & Farris, 1969), the bootstrap (Felsenstein, 1985), and randomisation and permutation tests (Archie, 1989; Faith & Cranston, 1991). More recently, Lewis et al. (2016) have sought to measure the information content of original data with trees generated through Bayesian analysis, suggesting a comparison of the entropy of the prior distribution with that of the posterior distribution. If they are identical, then the maximum amount of information from the data would be revealed in the structure of the tree.

Synthesis of information for answering different kinds of systematic questions

Use of different types and distillation of information in biological classification must relate to the kinds of questions being posed. The central question is obviously: What is the maximally informative classification for a particular group? To answer this question at the deepest level requires having answers to two other questions: What have been the processes of evolution that have operated within the group that have resulted in the data assembled in the basic data matrix, and, what has been the phylogeny of the group? In other words, for maximally predictive classification it is necessary to first understand the evolutionary mode of origin of a group (i.e. microevolution) as well as longer term evolutionary patterns resulting in phylogeny (i.e. macroevolution). Another important factor is the level of the taxonomic hierarchy at which the predictive classification will be formed. Information at the infraspecific level may not be useful at the interfamilial level, and vice versa.

The data in the basic data matrix are the way they are because of evolutionary processes of many types, and many different types of speciation have occurred during evolution of the angiosperms, especially progenitor-derivative (Crawford, 2010, 2014) and reticulate modes. Realisation of the complexity of these evolutionary origins mandates care in selection of characters and states for the basic data matrix to maximise final information content and to strengthen homologies. This deeper knowledge of origin of diversity allows greater precision in the collection and ordering of data to be used with the questions regarding phylogeny. Species known to have originated via progenitor-derivative processes (Crawford, 2010) cannot be interpreted as having had a branching origin from the ancestor. This type of speciation also occurs in peripheral geographic budding, oceanic island speciation, and polyploidy. Regarding reticulate modes, it has been estimated that all or nearly all of the current angiosperms have had polyploid origins (Soltis et al., 2009). This involves either allopolyploid mechanisms, which combine two genomes into a new lineage, or autopolyploidy, whereby doubling occurs within a single lineage. It would be no exaggeration to state that a large proportion of evolutionary processes in the flowering plants would be other than via dichotomous allopatric speciation. Most studies of speciation now employ some type of molecular data, often nucleotide sequences and/or population genetic markers such as AFLPs, or nuclear or organellar microsatellites, and now Next Generation Sequencing (NGS) techniques (e.g. Hörandl & Appelhans, 2015). It is appropriate once more to emphasise that to understand speciation requires first having a general understanding of relationships such as provided by a comprehensive revisionary study (Stuessy, 1975, 1993, 2011; Marhold & Stuessy, 2013). One can hardly study modes of speciation if there is no clear view of which species are closely related to each other. Here the historical information, inferences, and hypotheses that have been accumulated for a group become extremely important.

The second question that needs to be answered for a group is its phylogeny. Most investigations now require a minimum of nucleotide data from both the nucleus and chloroplast (and/or mitochondrion) and several sequences are preferred. With NGS methods, the amount of easily obtainable nucleotide data is becoming massive. The challenge now is to find ways of sorting through the literally millions of comparative base pairs for those that seem most diagnostic for revealing phylogeny. At this early stage, we simply do not have any community standards for such information syntheses. Allied with the new abundance of nucleotide data are new statistical methods for seeking comparative phylogenetic signal from within them, and to make these results interpretable to people in some sort of graphic display. Although traditionally such results have been synthesised in dendrograms, it is suspected that in the future we might find sophisticated mathematical modes of interpretation far beyond the simple tree-building approaches now in use.

The final step in information synthesis is the construction of the predictive classification. Here all accumulated data and inferences are marshalled for constructing the most information-rich hierarchical structure. The phylogenetic analysis is key here, and if done well, it should portray evolutionary relationships at the level of synthesis as best as can be done at present. The challenge is to utilise as much of the phylogenetic information as possible for purposes of classification. The branching dimension (cladistic relationship) is clearly significant, but this only gives one aspect of the total information. The degree of divergence among taxa (or lineages) is also most significant as this registers genetic and evolutionary change through time. There are several ways of measuring such divergence quantitatively, but it basically involves determining precisely the cohesiveness and distinctiveness of taxa and groups to each other (Stuessy, 2013). All evolutionary groups must come from common ancestors, i.e. they must be monophyletic, sensu lato, which involves holophyly and paraphyly (Ashlock, 1971; Stuessy, 2009a).

Conclusions

It should be obvious from the preceding discussion that to construct maximally predictive classification requires considerable biological understanding, assembly of data, and numerous quantitative distillations. One might hardly expect otherwise. The varied structural and reproductive diversity of organisms, the many different modes of speciation, the broad spectrum of available data, and the numerous algorithms for synthesis of information suggest many challenges in a complex process.

For understanding processes of evolution, such as population divergence, speciation, hybridisation, and polyploidy, there can be little doubt that molecular markers deriving from population genetics studies are most useful. These questions can only be resolved definitively at the genetic level. Sophisticated data analyses are needed, and simple tree-building algorithms are clearly inappropriate. Another way of saying this is that cladistic concepts and methods are unsuitable for population-level questions. This pattern of information tends to be mosaic and hence far from the reach of simple approaches that reveal only dichotomous patterns.

For questions relating to the reconstruction of phylogeny, nucleotide sequences are required for this level of information synthesis. What we seek is the most accurate representation of phylogeny possible usually as a dendrogram in two or three dimensions, recognising that future investigations may reveal methods we cannot at present envisage. It is at this level that NGS techniques and huge quantities of data will have maximum impact. Phylogeny is a fundamental basis for the construction of classification at the generic level and above. At this level of the hierarchy, the population genetic markers that are efficacious at the infraspecific and specific levels are no longer of value. Nucleotide sequences are now fundamental. One might argue that the process of extinction is now more significant than at the lower levels of the hierarchy, as this produces gaps between groups that help to define their distinctiveness.

For investigations dealing with adaptations, morphology remains central because it is the phenotype that interacts directly with the environment. Working with morphology is not easy, especially due to the challenges of defining characters and delimiting character states. Furthermore, structures of flowering plants can be extremely plastic, making selection of stable, genetically controlled features for investigation difficult. Nonetheless, for questions that deal with ecological factors at the specific and infraspecific level, morphology must be analysed, distilled, and synthesised. In evaluation of phylogeny, morphology can also be important for understanding innovations within lineages that have explosively radiated into particular ecological zones.

A further dimension regarding information in biological classification that needs attention is the tie to information theory (e.g. Shannon, 1948; Shannon & Weaver, 1949; Ash, 1965; Pierce, 1980; Kåhre, 2002; MacKay, 2003). From a general perspective, it would be hard to imagine that mathematical information theory would not offer something of importance to our understanding of information in biological classification. There is a clear parallel between the basic sequence of information communication and that of phylogeny construction. In the former the sequence (Ash, 1965) is from: source to encoder, to noisy channel, to decoder, and to destination. In the latter it is: the dynamics of the micro-processes of evolution over time, being encoded as the original true phylogeny, then receiving interference from reversals, parallelisms, reticulations, and extinctions, yielding the modified true phylogeny, and finally being decoded in construction of the phylogenetic tree. Much of information theory focuses on ways to use and manipulate bits of information, which nowadays falls neatly into the digital computer age. The challenges to biological classification are very much the same: how to conceptualise and delimit characters and states, how to evaluate them, and how to seek patterns in the data.

A few initial applications of information theory in classification have been attempted. One was by Duncan & Estabrook (1976), based on Estabrook (1971), whereby characters coded with multiple states were assumed to have more information than those with only two (binary) states. This measure was used successfully to evaluate the information content of different classifications of the *Ranunculus hispidus* Michx. complex (Ranunculaceae). The same measure was used by Carpenter (1993) to evaluate information contained within both cladistic and evolutionary (phyletic) classifications of fusilier fishes, in which more information for the latter was found. Another more recent mathematical contribution was by Craig & Stone (2015) who showed that as new apomorphic or synapomorphic characters were added to the data matrix, a cladogram gained in information content up to a certain limit. More studies of this nature are needed.

ACKNOWLEDGMENTS. Editorial help from Alistair Hay is greatly appreciated. This paper is dedicated, with deep admiration, to Professor David Mabberley on the occasion of his 70th birthday. He has made major contributions to understanding the history and evolution of the vascular plants.

References

Archie, J.W. (1989). A randomization test for phylogenetic information in systematic data. *Syst. Zool*. 38: 239–252.

Ash, R.B. (1965). *Information Theory*. New York: Interscience.

Ashlock, P.D. (1971). Monophyly and associated terms. *Syst. Zool*. 20: 63–69.

Baum, D.A. & Smith, S.D. (2012). *Tree Thinking: An Introduction to Phylogenetic Biology*. Greenwood Village, Colorado: Roberts & Co.

Brady, R.H. (1985). On the independence of systematics. *Cladistics* 1: 113–126.

Bramwell, D. (2002). How many plant species are there? *Pl. Talk* 28: 32–34.

Brown, C.M., Arbour, J.H. & Jackson, D.A. (2012). Testing of the effect of missing data estimation and distribution in morphometric multivariate data analyses. *Syst. Biol*. 61: 941–954.

Carpenter, K.E. (1993). Optimal cladistic and quantitative evolutionary classifications as illustrated by fusilier fishes (Teleostei: Caesionidae). *Syst. Biol*. 42: 142–154.

Craig, W. & Stone, J. (2015). Information and phylogenetic systematic analysis. *Information* 6: 1–x.

Crawford, D.J. (2010). Progenitor-derivative species pairs and plant speciation. *Taxon* 59: 1413–1423.

Crawford, D.J. (2014). Hybridization and homoploid hybrid speciation; polyploidy. In: Stuessy, T.F., Crawford, D.J., Soltis, D.E. & Soltis, P.S. *Plant Systematics: The Origin, Interpretation, and Ordering of Plant Biodiversity*, pp. 125–155. Königstein, Germany: Koeltz Scientific Books.

Crisci, J.V. & Stuessy, T.F. (1980). Determining primitive character states for phylogenetic reconstruction. *Syst. Bot.* 5: 112–135.

Davis, P.H. & Heywood, V.H. (1963). *Principles of Angiosperm Taxonomy*. Princeton, New Jersey: Van Nostrand.

Duncan, T. & Estabrook, G.F. (1976). An operational method for evolutionary classifications. *Syst. Bot.* 1: 373–382.

Estabrook, G.F. (1971). Some information theoretic optimality criteria for general classification. *Math. Geol.* 3: 203–207.

Faith, D.P. & Cranston, P.S. (1991). Could a cladogram this short have arisen by chance alone? On permutation tests for cladistic structure. *Cladistics* 7: 1–28.

Farris, J.S. (1969). On the cophenetic correlation coefficient. *Syst. Zool.* 18: 279–285.

Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* 39: 783–791.

García-Laencina, P.J., Sancho-Gómez, J.-L. & Figueiras-Vidal, A.R. (2010). Pattern classification with missing data: a review. *Neural Comput. & Applic*. 19: 263–282.

Govaerts, R. (2001). How many species of seed plants are there? *Taxon 50*: 1085–1090.

Govaerts, R. (2003). How many species of seed plants are there? – a response. *Taxon* 52: 583–584.

Grant, V. (1950). The protection of the ovules in flowering plants. *Evolution* 4: 179–201.

Hawkins, J.A. (2000). A survey of primary homology assessments: different botanists perceive and define characters in different ways. In: Scotland, R. & Pennington, R.T. (eds) *Homology and Systematics: Coding Characters for Phylogenetic Analysis*, pp. 22–53. London: Taylor & Francis.

Hennig, W. (1950). *Grundzüge einer Theorie der phylogenetischen Systematik*. Berlin: Deutscher Zentralverlag.

Hennig, W. (1966). *Phylogenetic Systematics*, transl. D.D. Davis and R. Zangerl. Urbana: University of Illinois Press.

Hörandl, E. & Appelhans, M.S. (eds) (2015). *Next-generation Sequencing in Plant Systematics*. Königstein, Germany: Koeltz Scientific Books.

IUCN. (2010). *Red List of Threatened Species. Summary statistics for globally threatened species; Table 1, Numbers of threatened species by major groups of organisms (1996–2013).* https://www.iucnredlist.org. Accessed 7 Jan. 2019.

Kåhre, J. (2002). *The Mathematical Theory of Information*. Dordrecht: Kluwer Academic Publishers.

Kemp, T.S. (1985). Models of diversity and phylogenetic reconstruction. *Oxford Surv. Evol. Biol.* 2: 135–158.

Kluge, A.G. & Farris, J.S. (1969). Quantitative phyletics and the evolution of anurans. *Syst. Zool.* 18: 1–32.

Lewis, P.O., Chen, M.-H., Kuo, L., Lewis, L.A., Fučíková, K., Neupane, S., Wang, Y.-B. & Shi, D. (2016). Estimating Bayesian phylogenetic information content. *Syst. Biol.* 65: 1009–1023.

MacKay, D.J.C. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge: Cambridge University Press.

Maddison, W.P. (1993). Missing data versus missing characters in phylogenetic analysis. *Syst. Biol.* 42: 576–581.

Marhold, C. & Stuessy, T.F. (eds) (2013). The future of botanical monography: Report from an international workshop, 12–16 March 2012, Smolenice, Slovak Republic. *Taxon* 62: 4–20.

Mickevich, M.F. & Platnick, N.I. (1989). On the information content of classifications. *Cladistics* 5: 33–47.

Miller, J. (1996). Collecting methodologies for plant samples for pharmaceutical research. In: Stuessy, T.F. & Sohmer, S.H. (eds) *Sampling the Green World: Innovative Concepts of Collection, Preservation, and Storage of Plant Diversity*, pp. 74–87. New York: Columbia University Press.

Mora, C., Tittensor, D.P., Adl, S., Simpson, G.B. & Worm, B. (2011). How many species are there on earth and in the ocean? *PLoS Biol*. 9(8): e1001127.

Moran, K., King, S.R. & Carlson, T.J. (2001). Biodiversity prospecting: Lessons and prospects. *Ann. Rev. Anthropol.* 30: 505–526.

Patterson, C. (ed.) (1987). *Molecules and Morphology in Evolution: Conflict or Compromise?* Cambridge: Cambridge University Press.

Patterson, C., Williams, D.M. & Humphries, C.J. (1993). Congruence between molecular and morphological phylogenies. *Annual Rev. Ecol. Syst*. 24: 153–188.

Pierce, J.R. (1980). *An Introduction to Information Theory: Symbols, Signals and Noise*, 2nd ed. New York: Dover.

Pietsch, T.W. (2012). *Trees of Life: A Visual History of Evolution*. Johns Hopkins University Press: Baltimore.

Platnick, N.I. (1985). Philosophy and the transformation of cladistics revisited. *Cladistics* 1: 87–94.

Reid, G. & Sidwell, K. (2002). Overlapping variables in botanical systematics. In: MacLeod, N. & Forey, P.L. (eds) *Morphology, Shape and Phylogeny*, pp. 53–66. London: Taylor & Francis.

Rohlf, F.J. (1974). Methods of comparing classifications. *Annual Rev. Ecol. Syst*. 5: 101–114.

Scotland, R.W. & Wortley, A.H. (2003). How many species of seed plants are there? *Taxon* 52: 101–104.

Scotland, R.W., Olmstead, R.G. & Bennett, J.R. (2003). Phylogeny reconstruction: The role of morphology. *Syst. Biol*. 52: 539–548.

Shannon, C.E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* 27: 379–423, 623–656.

Shannon, C. & Weaver, W. (1949). *The Mathematical Theory of Communication*. Urbana: University of Illinois Press.

Sneath, P.H.A. & Sokal, R.R. (1973). *Numerical Taxonomy: The Principles and Practice of Numerical Classification*. San Francisco: Freeman & Co.

Sokal, R.R. & Rohlf, F.J. (1962). The comparison of dendrograms by objective methods. *Taxon* 11: 33–40.

Sokal, R.R. & Sneath, P.H.A. (1963). *Principles of Numerical Taxonomy*. San Francisco: Freeman & Co.

Soltis, D.E. (2014). Congruence and consensus. In: Stuessy, T.F., Crawford, D.J., Soltis, D.E. & Soltis, P.S. *Plant Systematics: The Origin, Interpretation, and Ordering of Plant Biodiversity*, pp. 193–208. Königstein, Germany: Koeltz Scientific Books.

Soltis, D.E., Fishbein, M. & Kuzloff, R.K. (2003). Reevaluating the evolution of epigyny: Data from phylogenetics and floral ontogeny. *Int. J. Pl. Sci.* 164(5 Suppl.): S251–S264.

Soltis, D.E., Albert, V., Leebens-Mack, J., Bell, C.D., Paterson, A., Zheng, C., Sankoff, D., Wall, P.K. & Soltis, P.S. (2009). Polyploidy and angiosperm diversification. *Amer. J. Bot*. 96: 336–348.

Stevens, P.F. (1980). Evolutionary polarity of character states. *Annual Rev. Ecol. Syst*. 11: 333–358.

Stevens, P.F. (1991). Character states, morphological variation, and phylogenetic analysis: A review *Syst. Bot*. 16: 553–583.

Stuessy, T.F. (1975). The importance of revisionary studies in plant systematics. *Sida* 6: 104–113.

Stuessy, T.F. (1987). Explicit approaches for evolutionary classification. *Syst. Bot*. 12: 251–262.

Stuessy, T.F. (1990). *Plant Taxonomy: The Systematic Evaluation of Comparative Data*. New York: Columbia University Press.

Stuessy, T.F. (1993). The role of creative monography in the biodiversity crisis. *Taxon* 42: 313–321.

Stuessy, T.F. (1997). Classification: more than just branching patterns of evolution. *Aliso* 15: 113–124.

Stuessy, T.F. (2009a). *Plant Taxonomy: The Systematic Evaluation of Comparative Data*, 2nd ed. New York: Columbia University Press.

Stuessy, T.F. (2009b). Paradigms in biological classification (1707–2007): Has anything really changed? *Taxon* 58: 68–76.

Stuessy, T.F. (2011). Importance of the botanical monograph. In: Stuessy, T.F. & Lack, H.W. (eds) *Monographic Plant Systematics: Fundamental Assessment of Plant Biodiversity*, pp. 7–14. Ruggell: A.R.G. Gantner Verlag.

Stuessy, T.F. (2013). Schools of data analysis in systematics are converging, but differences remain with formal classification. *Taxon* 62: 876–885.

Stuessy, T.F. & König, C. (2008). Patrocladistic classification. *Taxon* 57: 594–601.

Stuessy, T.F., Crawford, D.J., Soltis, D.E. & Soltis, P.S. (2014a). *Plant Systematics: The Origin, Interpretation, and Ordering of Plant Biodiversity*. Königstein, Germany: Koeltz Scientific Books.

Stuessy, T.F., König, C. & López Sepúlveda, P. (2014b). Paraphyly and endemic genera of oceanic islands: Implications for conservation. *Ann. Missouri Bot. Gard*. 100: 50–78.

Systma, K.J. (1990). DNA and morphology: Inference of plant phylogeny. *Trends Ecol. Evol*. 5: 104–110.

Thorne, R.F. (2002). How many species of seed plants are there? *Taxon* 51: 511–522.

Voss, E. (1952). The history of keys and phylogenetic trees in systematic biology. *J. Sci. Lab. Denison Univ.* 43(1): 1–15.

Wiens, J.J. & Morrill, M.C. (2011). Missing data in phylogenetic analysis: reconciling results from simulations and empirical data. *Syst. Biol*. 60: 719–731.