9-1-2006

The Evolution of Organismal Complexity in Angiosperms as Measured by the Information Content of Taxonomic Descriptions

J. Gordon Burleigh

Justen B. Whittall
Santa Clara University, jwhittall@scu.edu

Michael J. Sanderson

The Evolution of Organismal Complexity in Angiosperms as Measured by the Information Content of Taxonomic Descriptions

J. Gordon Burleigh[1], Justen B. Whittall[1] and Michael J. Sanderson[1]

[1]Section of Evolution and Ecology; University of California, Davis, CA 95616
jgburleigh@ucdavis.edu

Abstract

We describe an information theoretic method for measuring relative organismal complexity. The complexity measure is based on the amount of information contained in formal taxonomic descriptions of organisms. We examine the utility of this measure for quantifying the complexity of plant families. The descriptions are subjective by nature, but we find a significant correlation in the complexity values of plant families from two independently authored sets of formal taxonomic descriptions. An analysis of the evolution of complexity across angiosperms found evidence of a pattern of increasing complexity. Our measure of complexity provides an operational definition of complexity that may be applied to any group of organisms and will enable further empirical studies of the evolution of complexity.

Introduction

While the evolution of biological complexity has interested scientists for many years, complexity has been notoriously difficult to define- let alone quantify (see reviews in Bonner, 1988; McShea, 1991; Maynard Smith and Szathmáry, 1995; Gould, 1996; Carroll, 2001; Adami, 2002). Several approaches to quantifying complexity have focused on measurements of a single, homologous trait, such as arthropod limbs (Cisne, 1974), mammalian vertebral columns (McShea, 1993), or septal sutures of ammonoids (Saunders et al., 1999). However, since the complexity in a single trait is not necessarily indicative of the complexity in all traits, these measurements may not reflect total organismal complexity. Furthermore, these comparisons rely on assumptions of homology, and they can only be applied to a limited number of organisms and traits. For example, if one is measuring complexity based on vertebral columns, it is impossible to compare the complexity of mammals and insects. Approaches to quantify whole organism complexity have included counting the number of cell types (e.g., Bonner, 1988; Valentine et al., 1994) or the number of descriptive terms for different groups of organisms (Schopf et al., 1975). Similarly, attempts to measure the functional complexity of organisms have used measures based on the number of morphological, behavioral, or physiological parts (McShea, 2000) and the number of levels in an organizational hierarchy in an organism (Nehaniv and Rhodes, 2000). With the increase in genomic sequence data, there has been much interest in measuring the complexity of genomes (Adami et al., 2000; Lynch and Conery, 2003), but it is unclear if there is a relationship between genomic and morphological or structural complexity (e.g., Szathmáry et al., 2001; Hahn and Wray, 2002; Stellwag, 2004).

In order to study the evolutionary patterns of complexity, it is necessary to have an operational method to quantify complexity across large groups of organisms. Such a measure can be used to address questions regarding possible directionality of the evolution of complexity and the evolutionary correlates of changes in complexity. We describe a method to measure morphological and structural complexity of an entire organism or group of organisms based on the information contained in formal taxonomic descriptions. This is an extension and refinement of an idea of Schopf et al. (1975) that the richness of terminology used to describe an organism is an indication of its complexity. This measure is intended to directly reflect the knowledge of the taxonomic authorities and, indirectly the accumulated knowledge of their entire discipline. We test the validity of this method and illustrate its utility for evolutionary studies of angiosperms by examining two independently authored sets of plant family descriptions.

Methods

Our method for quantifying organismal complexity defines complexity as the minimum information required to describe an organism (e.g., Papetin, 1980; Saunders and Ho, 1981). This definition is related to the information theoretic notion of Kolmogorov complexity (Kolmogorov, 1965) focusing on the minimal information in a description of an object, not the object itself. The relative complexity of organisms can be measured based on the information content in a set of formal descriptions of the organisms. We begin with ASCII text files containing formal taxonomic descriptions. The information in a text description file is related to the size of the file and the heterogeneity and randomness of the characters within the file. The information content of a file is estimated by measuring its size after it has been compressed using a standard text compression tool.

Sources. We demonstrated our method using two sets of plant family descriptions. Cronquist (1981) provided descriptions for 373 families of angiosperms (flowering plants), while Judd et al. (2002) provided descriptions of

161 major families of land plants. Note that these two works adopted divergent principles for recognizing taxonomic groups ("evolutionary taxonomy" versus "phylogenetic taxonomy"; see Judd et al., 2002), and therefore agreement between them likely reflects signal that rises above this background of methodological differences. Both sources contain a formal description of the plant families followed by a general, informal discussion of the family. The formal family descriptions in both sources followed a strict format of presenting observations of specific sets of characters. Only the formal descriptions of the families from each book were digitized using flatbed scanner and standard text recognition tools (OCR). We checked the accuracy of the OCR for each family and corrected the text when necessary. Each family description (uncompressed) was then saved as an ASCII text file.

Both sources contain not only descriptions of the traits shared by all members of a plant family but also descriptions of trait variation within each plant family. Thus, it includes a measure of the complexity of the family and the complexity within the family. It is possible that a family could be composed of very simple but very diverse organisms. In such a case, the complexity of the organisms in the family would be low, but the complexity of the family could appear very high due to the description of the variation among organisms within the family. In order to compare the complexity of plant families, we edited each family description to prune out any descriptions that related to the variation within the family. To do this we followed a precise editing protocol. First, we removed any adjectives that describe the frequency with which a trait appears in a family (e.g., *always*, *often*, *frequently*, *sometimes*). If such an adjective implies that a trait is rarely found in the family (e.g., *seldom*, *infrequently*, *rarely*), we also deleted the text describing the trait. For example, "*often trait X*" would be edited to *"trait X*", but "*seldom trait X*" would be removed entirely. If a description says "*trait X or trait Y*", we deleted the word "*or*" and the text describing one of the traits. We always kept the first trait listed unless the description stated that the second trait was more common. For example, "*trait X or trait Y*" would be edited to "*trait X*", but "*trait X or more often trait Y*" would be edited to "*trait Y*". If there was a range of numbers, we always took the first number, again unless it stated that another number was more frequent. So the text "*2-6 of trait X*" would be edited to "*2 of trait X*". We also deleted any taxonomic names in the descriptions. Preliminary tests showed that it was very repeatable (data not shown).

After the family descriptions were edited to remove within family variation, each file was compressed using the GNU utility gzip. The complexity value for the file is the size of the compressed file measured in bytes. The rationale for the compression step is that any lengthy text description will contain uninformative redundancy that will tend to inflate its apparent information content. Compression removes this to a degree, although even the asymptotically optimal LZ compression of GNU compress cannot guarantee to remove all redundancy. We used gzip only because it is a commonly used tool text compression, but we do not claim that this is the optimal tool to remove redundancy in the taxonomic descriptions. While we feel that some sort of compression is an important part of our method for measuring complexity further tests are needed to identify the most appropriate methods for compression.

Comparison of Two Sources. If our measure of complexity relates to a biological property of the plant families and not the influence of the authors, then we would expect the measures of complexity from the two independently authored sources will be correlated. However, a comparison of the complexity values for plant families from the two sources is not straightforward. Not only do the two sources contain different numbers of family descriptions (Cronquist, 1981; Judd et al., 2002), they reflect very different notions of plant families (see APG II, 2002; Judd et al., 2002). Even descriptions with the same family names in our two sources may describe different sets of taxa. Thus, we limited the comparison to 123 families in which we determined the classifications were consistent. We performed a model II regression analysis to compare the complexity scores from the two sources. This is appropriate when both variables being compared are random variables (Sokal and Rohlf, 1995). The significance of the correlation was tested with a permutation test with 999 permutations implemented in the program Model II (http://www.bio.umontreal.ca/Casgrain/en/labo/model-ii.html#ref).

Evolutionary Analyses. To determine if there are any evolutionary patterns of changing complexity across angiosperms, we examined the Judd et al. (2002) and Cronquist (1981) data sets in a phylogenetic context. In order to do this, family-level complexity values were assigned to terminal nodes in an angiosperm-wide phylogeny (Soltis et al., 2000). Most of the Judd et al. (2002) families reflect the same family assignments used in the Soltis et al. (2000) tree. However, the family assignments of Cronquist (1981) frequently did not match. Therefore, we matched genera assignments from Cronquist (1981) to genera sampled in the Soltis et al. (2000) tree. When no generic overlap was found, we used Mabberly (1987) to match the Cronquist (1981) genera to their modern equivalents in the Soltis et al. (2000) tree. When more than one genus was sampled from the same family in the Soltis et al. (2000) phylogeny, we pruned the tree to represent only a single complexity value for a family. Terminal taxa in the Soltis et al. (2000) tree without corresponding complexity values also were pruned from the tree.

We examined both data sets for evidence of directional changes in complexity through the history of angiosperms using ancestral state reconstructions. Ancestral complexity values were reconstructed using squared-change parsimony (Maddison, 1991) as implemented in a modified r8s program (Sanderson, 2003). We then examined the change in complexity between each ancestral and descendent node

throughout the tree. We measured the mean and median change in complexity between all ancestral and descendant nodes throughout the tree, and we counted the number of increases and decreases in complexity. To assess significance of these measurements, we compared the mean change in complexity and the total number of positive or negative changes in complexity with a null distribution obtained by analyses of 100 random reshufflings of the terminal complexity values.

Results

Description of Data. The complexity scores from Judd et al. (2002) were usually lower than those of Cronquist (1981), but the distribution of complexity from both sources generally appears as a bell curve (Figure 1). The tails of both distributions contain some large complexity values relative to the other scores (Figure 1).

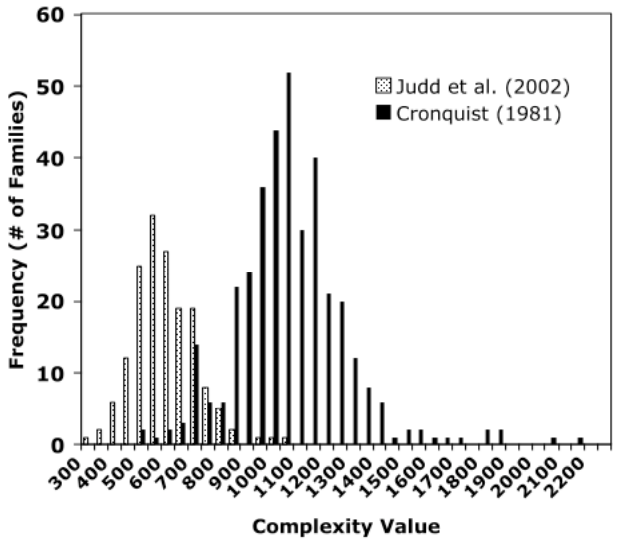
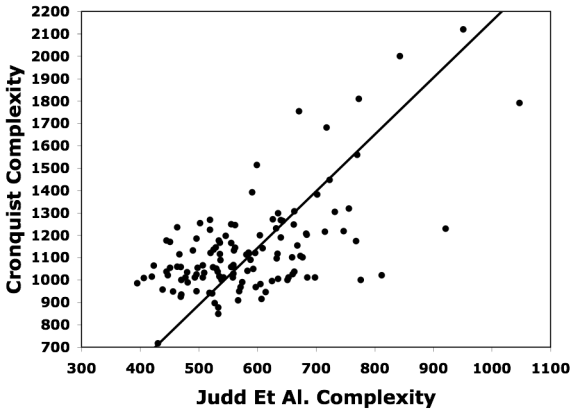


Figure 1. The distribution of plant family complexity values (in bytes) from Cronquist (1981) and Judd et al. (2002).

Correlation Among Complexity Measures From Different Sources. In the comparison of the complexity values from the 123 equivalent angiosperm families from Cronquist (1981) and Judd et al. (2002) for 123 plant families, the slope of the model II major axis regression line is 2.59 and the y-intercept is -380.88. In the model II ordinary least squares regression, the r^2 value is 0.38. A permutation test with 999 permutations indicates that the slopes of the major axis and ordinary least squares regression and the correlation coefficient were strongly significant (all $p \leq 0.001$).

Figure 2. Comparison of the complexity values in bytes of family descriptions from Judd et al. (2002) and Cronquist (1981). The complexity values are the size of the compressed files of the family descriptions. The line represents the model II major axis regression.



Patterns of Complexity Evolution Across Angiosperms.

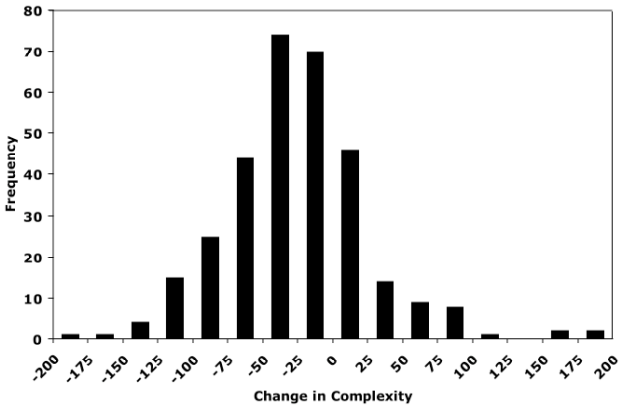
Table 1. The changes in complexity across all ancestor and descendent node comparisons based on the angiosperm phylogeny of Soltis et al. (2000). This table shows the mean and median complexity difference between ancestral and descendant nodes as well as the overall number of positive and negative changes in complexity throughout the tree. All complexity measures are in bytes.

	Cronquist (1981)	Judd et al. (2002)
Mean Difference	**2.18**	**-0.55**
Median Difference	**2.77**	**-2.18**
Increases	**413**	**152**
Decreases	**379**	**164**

While the evolutionary analyses of the complexity values from Judd et al. (2002) are equivocal regarding the evolution of complexity, the evolutionary analyses of the Cronquist (1981) complexity values show some evidence of a directional trend toward increasing complexity angiosperms (Table 1; Figure 2). The data from Judd et al. (2002) shows small decreases in the mean and median complexity throughout the tree (Table 1). However, the randomization tests indicate no significant trend in the evolution of complexity. The mean difference in complexity and the number of increases in complexity throughout the tree are not significantly less than expected if the complexity values were randomly assigned to terminals ($p = 0.35$ and 0.43 respectively). In the complexity analyses using the Cronquist (1981) data, there are significantly more increases in complexity throughout the tree than we would expect if the complexity values were randomly assigned to terminals ($p \leq 0.01$). However,

the mean change in complexity is only weakly significant ($p = 0.06$).

A.



B.

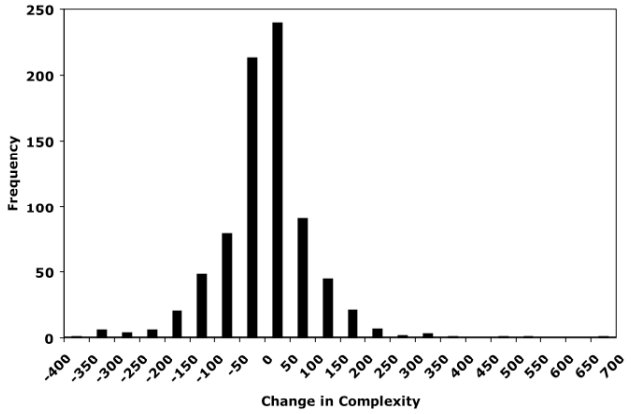


Figure 3. Frequency histogram illustrating the difference between ancestral and descendent complexity values in bytes throughout the phylogenetic tree. A) Judd et al. (2002) complexity values, B) Cronquist (1981) complexity values.

Discussion

We described a measure of complexity that is based on an information theoretic view of authoritative taxonomic descriptions. An obvious criticism of our method is that the size of the descriptions may reflect the level of familiarity or interest of the author in an organism rather than the overall complexity of the organism. It is easy to find instances in which this is the case. For example, while we confidently presume that all angiosperms have chromosomes, only some of the family descriptions from Cronquist (1981) contain chromosome counts. Thus, some of the differences in complexity among family descriptions from Cronquist (1981) may be due to the availability of data regarding chromosome numbers. Problems such as this may arise more frequently in Cronquist (1981), who attempted an exhaustive description of all angiosperm families rather than in Judd et al. (2002), which is limited to shorter descriptions of a smaller number of "major" plant families. Still, both sources followed a strict format for describing sets of traits from each plant family. Also, if the differences in this complexity measure reflect the interest of the author or the amount of information available for plant families, this is unlikely to introduce a bias to analyses of the evolution of complexity. They would only affect analyses if there is a phylogenetic structure to the biases of information. More likely, these differences only will add noise to analyses of evolutionary complexity. Using taxonomic descriptions to quantify complexity also implicitly assumes that the traits that are described are those that contribute to the complexity of organisms. The taxonomic descriptions tend to focus on traits that can be used to distinguish among organisms, and we might assume that there is less variation in the traits that are not described.

The closest precedent to our method for measuring organismal complexity is the use of the number of descriptive terms of an organism or group of organisms as a measure of complexity by Schopf et al. (1975). This has been criticized for the potential effect of observer bias (McShea, 1990). Yet there are reasons to believe that our method might be less subject to observer biases. First, these descriptions were written with no apparent interest or regard for examining the complexity of the organisms. Also, other measures have been criticized for exhibiting an obvious trend before analysis (McShea, 1993), and such a trend is not obvious from the plant family descriptions. Even if the authors did introduce their personal biases regarding the complexity of the organism into the description, the evolutionary analyses are based on phylogenies built by a separate set of authors. Counting words seems like an imperfect method for quantifying complexity, since different words may relate to objects or descriptions that vary greatly in importance. However, there is a similar downside to compressing entire description files. In our method, word length is related to complexity. For example, the words "dark orange" would represent more complexity than the word "red" though they have similar meanings.

Any description of an organism or group of organisms contains only a small selection of all possible observations, and thus it will depend greatly on the observer (e.g., Saunders and Ho, 1981). Thus, our measure of complexity is inherently subjective. Since the observations and styles of different authors may differ greatly, it may be difficult to compare complexity values from different authors directly. For example, the family descriptions from Judd et al. (2002) generally have smaller complexity values than equivalent family descriptions from Cronquist (1981; Figures 1 and 2). Yet this does not mean that the complexity values do not reflect the same relative complexity. If they do reflect the relative complexity of the plant families, then there should be a correlation between the information content from different authors even if the raw complexity values differ. Thus, finding a correlation between the relative information content in

plant families of Cronquist (1981) and Judd et al. (2002) is an important step in validating our method for measuring complexity. We suggest that it will be important to confirm patterns in the evolution of complexity based on our method by examining multiple, independently authored sources. The validity and utility of our measure of complexity may best be demonstrated by finding significant trends or correlations concerning the evolution of complexity. While this assumes that such trends exist, finding a significant trend would be highly unlikely if the complexity values were meaningless or obscured by random error.

The analyses of the complexity measurements from Cronquist (1981) indicate some evidence of increasing complexity throughout the history of angiosperms. This is consistent with many historical expectations as well as some empirical studies suggesting that morphological and structural complexity increase through time (see Bonner, 1988; McShea, 1991, 2001; Valentine et al., 1994; Adami, 2002; but see McShea, 1993; Gould, 1996). This result is most evident in the overall number of increases in complexity from ancestor to descendent nodes in the phylogeny, and we would not expect it to be due to the few very large complexity values in the Cronquist (1981) data set (Figure 1). It appears to be due to an excess of small changes in complexity (Figure 2b). Still, the Cronquist (1981) data suggests there are several large shifts in complexity throughout the evolution of angiosperms (Figure 3b). The three largest shifts in complexity are in branches leading to the Asteraceae, Poaceae, and Orchidaceae (not shown). These are also among the largest and most studied angiosperm families. It is possible that the high complexity values for these families results from the amount of attention they have received from botanists rather than their inherent complexity. However, the flowers and floral structures from all three families intuitively seem very complex. The high complexity of these families also raises the possibility that increases in complexity are correlated with increased diversification, and it will be interesting to test this hypothesis throughout the full angiosperm tree.

While the apparent trend for increasing complexity in the Cronquist (1981) data is very intriguing, it will be important to test this data more thoroughly. This may include using evolutionarily meaningful branch lengths in the ancestral state reconstructions, performing significance tests with evolutionary simulations of traits underlying the complexity scores, and incorporating uncertainty in the tree topology. It also will be important to further examine the data set from Judd et al. (2002) to determine why it shows no evidence of increasing complexity. The Judd et al. (2002) data set contains fewer descriptions, and the descriptions are smaller than those of Cronquist (1981; Figure 1). It is possible the tests lack the power to detect a subtle trend in the evolution of complexity from the Judd et al. (2002) data. Since the data from Cronquist (1981) and Judd et al. (2002) cover different sets of families, it also is possible that there really is no trend in the evolution of complexity in the families it examines.

The results of our analyses motivate further tests using this measure of complexity. The validity of the complexity measure may be tested further by examining complexity scores based on other texts and in other systems, especially those in which there is previous evidence for increasing complexity. It will be informative to compare our measure of complexity with data from other measures of complexity. One advantage of our measure of complexity is that it describes organismal complexity with a single continuous variable that is very amenable to evolutionary analyses. Thus, it will be simple to use in phylogenetically informed tests to examine the effects of complexity on other traits (e.g., Felsenstein, 1985). For example, it is possible to use this measure to explicitly test hypotheses regarding the evolution of complexity with respect to changes in environment, life history, or diversification rate. It also will be interesting to compare our measure of organismal complexity with some measure of genomic complexity.

The wealth of new genomic data has generated much interest in measuring genomic complexity, sometimes using information theoretic approaches (e.g., Adami et al., 2000). Digital library projects promise to make accessible an equally rich domain of data in the form of digital descriptions of organisms in the not too distant future, drawing from hundreds of years of detailed biological observations. Though our measure of complexity is not without potential flaws, we hope that this study will inspire new discussion not only about quantifying evolutionary patterns of changes in complexity but also about utilizing data from digital library projects for evolutionary studies.

Acknowledgements

This work was supported by NSF grant No. 0431154 and a UC Davis postdoctoral fellowship in comparative biology to JBW.

References

Adami, C. 2002. What is complexity? BioEssays 24: 1085-1094.

Adami, C.; Ofria, C.; Collier, T. C. 2000. Evolution of biological complexity. Proceedings of the National Academy of Sciences, U.S.A. 97:4463-4468.

APG II (Angiosperm Phylogeny Group). 2003. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants. APG II. Botanical Journal of the Linnean Society 141:399-436.

Bonner, J. T. 1988. The evolution of complexity by means of natural selection. Princeton, NJ: Princeton University Press.

Carroll, S. B. 2001. Chance and necessity: the evolution of morphological complexity and diversity. Nature 409: 1102-1109.

Cisne, J. L. 1974. Evolution of the world fauna of aquatic free-living arthropods. Evolution 28: 337-363.

Cronquist, A. 1981. An integrated system of classification of flowering plants. New York, NY: Columbia University Press.

Felsenstein, J. 1985. Phylogenies and the comparative method. American Naturalist 125: 1-15.

Gould, S. J. 1996. Full house. New York: Harmony Books.

Judd, W. S.; Campbell, C. S.; Kellogg, E. A.; Stevens, P. F.; Donoghue, M. J. 2002. Plant systematics: a phylogenetic approach. Second edition. Sunderland, MA: Sinauer Associates.

Kolmogorov, A. N. 1965. Three approaches to the quantitative definition of information. Problems of information transmission 1: 4-7.

Lynch, M.; Conery, J. S. 2003. The origins of genome complexity. Science 302: 1401-1404.

Mabberley, D. J. 1987. The plant-book: a portable dictionary of the higher plants. Cambridge: Cambridge University Press.

Maddison, W. P. 1991. Squared-change parsimony reconstructions of ancestral states for continuous-valued characters on a phylogenetic tree. Systematic Zoology 40: 304-314.

Maynard Smith, J.; Szathmáry, E. 1995. The major transitions in evolution. New York: Freeman.

McShea, D. W. 1991. Complexity and evolution: what everybody knows. Biology and Philosophy 6:303-324.

McShea, D. W. 1993. Evolutionary change in the morphological complexity of the mammalian vertebral column. Evolution 47:730-740.

McShea, D. W. 2000. Functional complexity in organisms: parts as proxies. Biology and Philosophy 15:641-668.

McShea, D. W. 2001. The hierarchical structure of organisms: a scale and documentation of a trend in the maximum. Paleobiology 27:405-423.

Nehaniv, C. L.; Rhodes, J. L. 2000. The evolution and understanding of hierarchical complexity in biology from an algebraic perspective. Artificial Life 6: 45-67.

Pepetin, F. 1980. On order and complexity. I. General considerations. Journal of Theoretical Biology 87: 421-457.

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

Saunders, P. T.; Ho, M.-W. 1981. On the increase of complexity in evolution II. The relativity of complexity and the principle of minimum increase. Journal of Theoretical Biology 90: 515-530.

Saunders, W. B.; Work, D. M.; Nikolaeva, S. V. 1999. Evolution of complexity in Paleozoic ammonoid sutures. Science 286: 760-763.

Schopf, T. J. M.; Raup, D. M.; Gould, S. J.; Simberloff, D. S. 1975. Genomic versus morphologic rates of evolution: influence of morphologic complexity. Paleobiology 1: 63-70.

Sokal, R. R.; Rohlf, F. J. 1995. Biometry – the principles and practice of statistics in biological research. 3rd edition. New York: W. H. Freeman.

Soltis, D. E.; Soltis, P. S.; Chase, M. W.; Mort, M. E.; Albach, D. C.; Zanis, M.; Savolainen, V.; Hahn, W. H.; Hoot, S. B.; Fay, M. F.; Axtell, M.; Swensen, S. M.; Prince, L. M.; Kress, W. J.; Nixon, K. C.; Farris, J. S. 2000. Angiosperm phylogeny inferred from 18S rDNA, *rbcL*, and *atpB* sequences. Botanical Journal of the Linnean Society 133: 381-461.

Stellwag, E. J. 2004. Are genome evolution, organism complexity and species diversity linked? Integrative Computational Biology 44: 358-365.

Szathmáry, E.; Jordán, F.; Pál, C. 2001. Can genes explain biological complexity? Science 292: 1315-1316.

Valentine, J. W.; Collins, A. G.; Meyer, C. P. 1994. Morphological complexity increase in metazoans. Paleobiology 20: 131-142.