WILEY-BLACKWELL

Cladistics 26 (2010) 98–102

Cladistics

10.1111/j.1096-0031.2009.00270.x

A formula for maximum possible steps in multistate characters: isolating matrix parameter effects on measures of evolutionary convergence

Jennifer F. Hoyal Cuthill[a,b]*, Simon J. Braddy[b] and Philip C. J. Donoghue[b]

[a]*Department of Earth Sciences, University of Cambridge, Cambridge CB2 3EQ, UK;* [b]*Department of Earth Sciences, University of Bristol, Bristol BS8 1RJ, UK*

Accepted 14 May 2009

Abstract

To identify a biological signal in the distribution of homoplasy, it is first necessary to isolate non-biological factors affecting its measurement. The number of states per character in a phylogenetic data matrix may indicate evolutionary flexibility and, consequently, the likelihood of recurrent evolution. However, we show here that the number of states per character limits the maximum number of steps that may be inferred using parsimony. A formula is provided for the maximum number of steps that may be taken by a character with a given number of states and taxa. We show that as more character states are included the maximum proportion of steps that can be attributed to homoplasy falls, and the greatest amount of homoplasy measurable with the consistency index declines.

Homoplasy is the phylogenetic recurrence of cladistic characters (Archie, 1996). Homoplasy metrics provide universal measures of recurrent evolution, comparable between character and taxon samples far wider than those included in most geometric or multivariate statistical analyses of convergent evolution. As such, they facilitate quantitative analysis of a phenomenon variously described as the strongest evidence for evolutionary optimization (Conway Morris, 2003), a sign of developmental constraint on evolutionary possibilities (Wake, 1991), and the greatest obstacle on the path to a statistically consistent phylogenetic method (Foley, 1993). Unfortunately, these metrics do not simply express the consistency of character state change as they are subject to biases built-in by the parameters of the character matrix (Sanderson and Donoghue, 1989; Archie, 1996; Brooks, 1996). In order to uncover any biological signal underlying patterns in the incidence of convergence, it is therefore necessary to identify non-biological factors determining measures of character consistency such as the consistency index (*ci*, of Kluge and Farris (1969)). Well-studied matrix parameters known to affect homoplasy include the number of taxa (Archie, 1989a; Sanderson and Donoghue, 1989, 1996; Klassen et al., 1991; Hauser and Boyajian, 1997), number of characters (Archie, 1989b; Lamboy, 1994; Givnish and Sytsma, 1997a), and amount or rate of evolution (Felsenstein, 1978; Hauser and Boyajian, 1997; Simmons et al., 2004). Another matrix parameter deserving further attention is the number of character states. The number of states per character coded in a phylogenetic data matrix is of biological interest as this may be an indicator of the evolutionary flexibility of a lineage (Ricklefs and Renner, 2000) and, accordingly, of the likelihood of recurrent evolution (Chapman et al., 1979; Lamboy, 1994; Naylor and Kraus, 1995; Donoghue and Ree, 2000; Simmons et al., 2004). The commonly used multistate coding method preserves biological relationships between character states (Strong and Lipscomb, 1999), thereby recording morphological

*Corresponding author:
E-mail address: jfh41@cam.ac.uk

variability in the number of states per character. However, Naylor and Kraus (1995) noted that, among their randomly generated character data, the maximum number of extra steps achievable declined asymptotically as more states per character were allowed. This suggests that the number of states per character can act to limit maximum extra steps. The number of extra steps is the most direct measure of homoplasy and determines the minimum value that may be achieved by the consistency index, the simplest comparative homoplasy index. The relationship between the number of states per character and the amount of homoplasy measurable therefore represents a, less studied, factor potentially cross-study precluding comparison of homoplasy levels. Here we provide and prove a formula to calculate the maximum bound on the number of steps possible for any unordered character, providing the basis for more meaningful comparisons of the prevalence of convergence between studies.

Maximum possible steps

A character must take at least one step to explain the evolution of each derived state. Therefore, the minimum number of steps (m) required to explain the distribution of states without homoplasy is the number of character states minus one (e.g. Archie, 1996). Each homoplastic character state change, due to reversal from a derived to an ancestral state or to parallel evolution of a derived state on two branches of a tree, requires one extra step beyond the minimum. For a binary character, the maximum number of steps (g) that may be taken over any minimum length tree is equal to the frequency of the least frequent state (Meacham, 1981). For binary characters, this value is given by min (n_1, $t-n_1$) where n_1 is the number of taxa with state 1 and t is the number of taxa (Farris, 1973; Archie, 1989a). Similarly, for an unordered multistate character with n states, the maximum number of steps that may be taken over any minimum-length tree will equal the sum of the frequencies of the $n-1$ least frequent states. The highest number of steps possible for a character is then $g = t-F$, where F is the number of taxa with any one most frequent state (Steel and Penny, 2006). The total number of taxa with a minority state will be maximal for the most even distribution of states possible. Under this most even state distribution, $t-F$ is equal to the maximum possible number of steps that the character may show on any tree ($g_{\max}$).

$$g_{\max}(t,n) = t - \lceil t/n \rceil \quad (1)$$

where $\lceil t/n \rceil$ = the smallest integer $\geq t/n$, and is equal to the lowest possible number of taxa with any one most frequent state ($F_{\min}$).

Proof

The proof relies on the Erdös–Székely path system theorem (Erdös and Székely, 1992). This states that, for a character χ on a set of leaves X of a phylogenetic X-tree $\tau = (T; \phi)$, the most parsimonious length of χ on τ, $l(\chi, \tau)$, equals the maximum size of an Erdös–Székely path system for χ on τ (Erdös and Székely, 1992; Semple and Steel, 2003). For a character χ on τ, a set of directed paths P between the leaves of τ is an Erdös–Székely path system if the following conditions are met. (i) For each path p in P, the leaves at the tail and head of p ($\phi(x)$ and $\phi(y)$, respectively) do not have the same state ($\chi(x) \neq \chi(y)$) and p is therefore a proper path. (ii) If any two paths p_i and p_j in P share a branch in τ, then p_i and p_j travel in the same direction. (iii) If any two paths p_i and p_j in P share a branch in τ, then the end leaves of p_i and p_j ($\phi(x)$ and $\phi(y)$, respectively) do not have the same state ($\chi(x) \neq \chi(y)$).

The set X of t leaves can be divided into $F_{\min} = \lceil t/n \rceil$ groups, each containing one leaf with the majority state $\chi(F_i)$ and a set of leaves with minority states, among which each minority state $\chi(f_i)$ is represented not more than once. If two or more states have equal and maximal frequencies, any one can be designated the majority state (and all others minority states) without loss of generality. The leaves within each group $V_1, V_2, ..., V_{F_{\min}}$ can then be connected to form a binary phylogenetic tree τ_i. Let there be a set of paths P_i within each tree τ_i, within which each path connects the majority state leaf $\chi(F_i)$ to one and only one of the minority state leaves $\chi(f_i)$. (i) Let each path p_i in P_i be directed to travel in the same direction, with the majority state leaf $\chi(F_i)$ at the start of all paths. Then, the union of all the path sets ($\cup P_1, P_2, ..., P_{F_{\min}}$) is a set P of paths for which: (ii) each path is a proper path ($\chi(F_i) \neq \chi(f_i)$). (iii) No two paths that share a branch end in leaves of the same state ($\chi(f_i) \neq \chi(f_j)$). Therefore, the $F_{\min}$ binary phylogenetic trees $\tau_1, \tau_2, ..., \tau_{F_{\min}}$ can be connected to each other to form one binary phylogentic tree τ containing a set P of $t - \lceil t/n \rceil$ paths which is an Erdös–Székely path system for χ on τ. Thus, $t - \lceil t/n \rceil$ is a lower bound for $g_{\max}$ and there exists at least one binary phylogenetic tree τ for which the most parsimonious length of at least one character χ on τ $l(\chi, \tau) = g_{\max}(t,n) = t - \lceil t/n \rceil$. Given t leaves and n character states, it is clear that no tree τ can contain more than $t - \lceil t/n \rceil$ proper paths that form an Erdös–Székely path system, and therefore this is also an upper bound on $l(\chi, \tau)$ and $g_{\max}(t,n) = t - \lceil t/n \rceil$, completing the proof.

Special cases

Binary characters

Mickevich's (1978) formula, describing the relationship between the maximum possible number of steps

and the number of taxa for a binary character, is a special case of the general formula (1) where $n = 2$.

For even numbers of taxa and a binary character, $n = 2$ divides t to give an integer. Equation 1 then equates to

$$g_{\max}(t_{\text{even}}, 2) = t/2. \quad \text{(Mickevich, 1978)}$$

For odd numbers of taxa and a binary character, 1 must be subtracted from t before $n = 2$ divides t. From equation 1, we have

$$g_{\max}(t_{\text{odd}}, 2) = t - \frac{t+1}{2}$$

which gives

$$g_{\max}(t_{\text{odd}}, 2) = \frac{t-1}{2} \quad \text{(Mickevich, 1978)}$$

When states divides taxa

Steel and Penny's (2006) formula, giving maximum extra steps ($h_{\max}$) when the number of states divides into the number of taxa (to give an integer), reduces to a special case of equation 1. From equation 1, we have

$$g_{\max}(n|t) = t - t/n.$$

This equates to

$$g_{\max}(n|t) = t\left(1 - \frac{1}{n}\right).$$

The maximum extra steps for an individual character is $h_{\max} = g_{\max} - m$. As the minimum steps a character may take is $m = n - 1$, the maximum possible number of extra steps is then

$$h_{\max}(n|t) = t\left(1 - \frac{1}{n}\right) - n + 1. \quad \text{(Steel and Penny, 2006)}$$

The minimum possible consistency index

The character consistency index equals the minimum number of steps divided by the actual number of steps (s) implied by a tree. This is the proportion of steps explained by homology (Farris, 1989). The lowest value that the *ci* may take, and the greatest amount of homoplasy measurable, on any minimal length tree will equal the minimum number of steps required divided by the maximum number of steps possible. The maximum number of steps possible on any tree for the whole character set ($G_{\max}$) is the sum of the individual character $g_{\max}$ values. However, it should be noted that $G_{\max}$ will only be achieved on a bush phylogeny (Farris, 1973; Mickevich, 1978; Archie, 1989a; Archie and Felsenstein, 1993). This is because the length of a binary phylogenetic tree depends on the degree of congruence between characters as well as the maximum number of steps each may show (g). Consequently, S will in practice always fall below $G_{\max}$ on binary trees (Archie and Felsenstein, 1993).

For a given number of taxa, the difference ($h_{\max}$) between the minimum (m) and maximum number of steps ($g_{\max}$) possible for any one character shows an initial increase (or plateau) and then declines, as the number of states increases (Fig. 1c), an effect similar to that observed among simulated character data by Naylor and Kraus (1995). Meanwhile, the minimum number of steps rises linearly with states per character (Fig. 1a). As a result, the proportion of the maximum possible number of steps (Fig. 1b) due to homology is lower for characters with fewer states. Equivalently, the minimum consistency index that can be obtained (Fig. 1d) will rise as states are added and the maximum proportion of steps that can be attributed to homoplasy will fall.

It follows that if one character has a greater number of states than another, its *ci* may also be higher even if it shows an equal or greater number of extra steps. For example, compare two most even character state distributions among eight taxa, one with two states and one with four. For the binary character, at least one step is required to explain the data without homoplasy and a maximum of four steps are possible. The four-state character requires a minimum of three steps and allows a maximum of six. If the binary character shows two steps, we have one extra step and a *ci* of 0.5. If the four-state character shows five steps we have two extra steps but a *ci* of 0.6.

Figure 1 illustrates the relationship between states and maximum number of steps for the most even distribution of states among taxa. However, less even state distributions allow fewer steps (Sanderson and Donoghue, 1989; Archie, 1996). If the normality of state distributions is constant with respect to the number of states (where we have no reason to expect less even state distributions among characters with higher or lower numbers of states), the positive correlation between the minimum *ci* and states per character will hold among less even state distributions, but the lower bar on the *ci* will be raised so that even less homoplasy is detectable. A probabilistic decline in the amount of homoplasy measured by the consistency index as its minimum achievable value increases is consistent with the results from the simulations of Naylor and Kraus (1995). The magnitude of the effect of number of states on maximum steps, and therefore the minimum *ci*, will depend on the number of taxa present. As the number of taxa increases, the gradient of the relationship between the number of states and the minimum consistency index declines and the lowest value achievable by the consistency index falls. Therefore, minimum possible

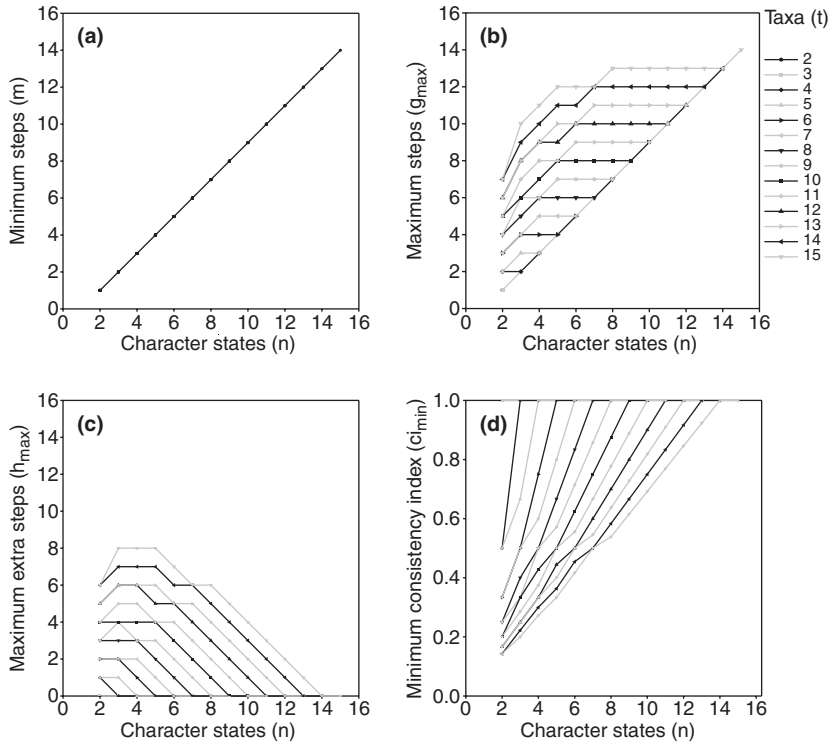

Fig. 1. (a) The minimum number of steps (m) possible for a character with a given number of states (n). For a character with n states and t taxa; (b) the maximum number of steps possible (g_{max}); (c) the maximum number of extra, homoplastic steps possible (h_{max}); and (d) the minimum value that the character consistency index may take (ci_{min}).

consistency index values for smaller datasets can be expected to be more strongly limited by the number of character states present.

A link between evolutionary flexibility and homoplasy has been suggested by several authors (e.g. Chapman et al., 1979; Lamboy, 1994; Naylor and Kraus, 1995; Donoghue and Ree, 2000; Simmons et al., 2004). However, evidence for a negative correlation between states per character and homoplasy (among character data in which the number of steps was free to vary) has been based on evolutionary simulations employing the consistency index as the homoplasy metric (Lamboy, 1994; Donoghue and Ree, 2000). The *ci* distributions generated fall within a range of values for which change in the amount of homoplasy that may be measured, rather than a decrease in the number of homoplastic state changes, presents an alternative explanation for the negative correlation observed. However, an increase in the number of homoplastic character state appearances when evolutionary options are strongly limited is a fundamental prediction of both functional optimization and developmental constraint hypotheses (Wake, 1991). We show here that, on minimum length trees, the number of states per character presents a maximum bound on character steps and determines the greatest amount of homoplasy that may be measured using the consistency index, the most direct comparative homoplasy metric. Comparative analyses of homoplasy and its relationship to morphological variability might proceed by: (i) assessing the tree length and its maximum bound as calculated from the data (G) against the theoretical maximum described here (G_{max}), (ii) comparing the battery of available homoplasy metrics in light of their sensitivities to the different matrix parameters (for a review see Archie, 1996), and (iii) treating character matrix parameters (e.g. Klassen et al., 1991; Givnish and Sytsma, 1997b), including the number of states per character, as covariates in multivariate statistical analyses of the distribution of homoplasy.

Acknowledgements

This manuscript is based on J. Hoyal Cuthill's MSc thesis prepared at the University of Bristol. S. Braddy and P. Donoghue were the project supervisors. The manuscript was written whilst working towards a PhD at the University of Cambridge supervised by S. Conway Morris and funded by a John Templeton Foundation studentship. We thank two anonymous reviewers for their highly constructive comments.

References

Archie, J.W., 1989a. Homoplasy excess ratios: new indices for measuring levels of homoplasy in phylogenetic systems and a critique of the consistency index. Syst. Zool. 38, 253–269.

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

Archie, J.W., 1996. Measures of homoplasy. In: Sanderson, M.J., Hufford, L. (Eds.), Homoplasy: The Recurrence of Similarity in Evolution. Academic Press, San Diego, pp. 153–188.

Archie, J.W., Felsenstein, J., 1993. The number of evolutionary steps on random and minimum length trees for random evolutionary data. Theor. Popul. Biol. 43, 52–79.

Brooks, D.R., 1996. Explanations of homoplasy at different levels of biological organization. In: Sanderson, M.J., Hufford, L. (Eds.), Homoplasy: The Recurrence of Similarity in Evolution. Academic Press, San Diego, pp. 6–34.

Chapman, R.W., Avise, J.C., Asmussen, M.A., 1979. Character space restrictions and boundary conditions in the evolution of multistate characters. J. Theor. Biol. 80, 51–64.

Conway Morris, S., 2003. Life's Solution: Inevitable Humans in a Lonely Universe. Cambridge University Press, Cambridge.

Donoghue, M.J., Ree, R.H., 2000. Homoplasy and developmental constraint: a model and example from plants. Am. Zool. 40, 759–769.

Erdös, P.L., Székely, L.A., 1992. Evolutionary trees: an integer multicommodity max-flow-min-cut theorem. Adv. Appl. Math. 13, 375–389.

Farris, J.S., 1973. On comparing the shapes of taxonomic trees. Syst. Zool. 22, 50–54.

Farris, J.S., 1989. The retention index and the rescaled consistency index. Cladistics 5, 417–419.

Felsenstein, J., 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27, 401–410.

Foley, R., 1993. Striking parallels in early hominid evolution. Trends Ecol. Evol. 8, 196–197.

Givnish, T.J., Sytsma, K.J., 1997a. Consistency, characters, and the likelihood of correct phylogenetic inference. Mol. Phylogenet. Evol. 7, 320–330.

Givnish, T.J., Sytsma, K.J., 1997b. Homoplasy in molecular vs. morphological data: the likelihood of correct phylogenetic inference. In: Givnish, T.J., Sytsma, K.J. (Eds.), Molecular Evolution and Adaptive Radiation. Cambridge University Press, Cambridge, pp. 55–102.

Hauser, D.L., Boyajian, G., 1997. Proportional change and patterns of homoplasy: Sanderson and Donoghue revisited. Cladistics 13, 97–100.

Klassen, G.J., Mooi, R.D., Locke, A., 1991. Consistency indices and random data. Syst. Zool. 40, 446–457.

Kluge, A., Farris, J.S., 1969. Quantitative phyletics and the evolution of anurans. Syst. Zool. 18, 1–32.

Lamboy, W.F., 1994. The accuracy of the maximum parsimony method for phylogenetic reconstruction with morphological characters. Syst. Bot. 19, 489–505.

Meacham, C.A., 1981. A probability measure for character compatibility. Math. Biosci. 57, 1–18.

Mickevich, M.F., 1978. Taxonomic congruence. Syst. Zool. 27, 143–158.

Naylor, G., Kraus, F., 1995. The relationship between s and m and the retention index. Syst. Biol. 44, 559–562.

Ricklefs, R.E., Renner, S.S., 2000. Evolutionary flexibility and flowering plant diversity: a comment on Dodd, Silvertown, and Chase. Evolution 54, 1061–1065.

Sanderson, M.J., Donoghue, M.J., 1989. Patterns of variation in levels of homoplasy. Evolution 43, 1781–1795.

Sanderson, M.J., Donoghue, M.J., 1996. The relationship between homoplasy and confidence in a phylogenetic tree. In: Sanderson, M.J., Hufford, L. (Eds.), Homoplasy: The Recurrence of Similarity in Evolution. Academic Press, San Diego, pp. 67–90.

Semple, C., Steel, M., 2003. Phylogenetics. Oxford University press, Oxford.

Simmons, M.P., Reeves, A., Davis, J.I., 2004. Character-state space versus rate of evolution in phylogenetic inference. Cladistics 20, 191–204.

Steel, M., Penny, D., 2006. Maximum parsimony and the phylogenetic information in multistate characters. In: Albert, V.A. (Ed), Parsimony, Phylogeny, and Genomics. Oxford University Press, Oxford, pp. 163–180.

Strong, E.E., Lipscomb, D., 1999. Character coding and inapplicable data. Cladistics 15, 363–371.

Wake, D.B., 1991. Homoplasy: The result of natural selection, or evidence of design limitations? Am. Nat. 138, 543–567.