Cladistics (1989)5:417–419

THE RETENTION INDEX AND THE RESCALED CONSISTENCY INDEX

James S. Farris[1]

[1] *Department of Ecology and Evolution, State University of New York, Stony Brook, NY 11794, U.S.A. and Department of Entomology, American Museum of Natural History, Central Park West at 79th Street, New York, NY 10024, U.S.A. and Department of Biology, City College, City University of New York, U.S.A.*

The consistency index, introduced by Kluge and Farris (1969) as a measure of fit of a character to a tree, has been widely and successfully employed, but might be capable of some improvement for certain applications. The purpose of this note is to describe two new indices, already in use in Hennig86, and to explain their interpretation.

The consistency index, c, is defined to be

$$c = m/s.$$

Here s denotes the amount of change in the character (for an integral character, number of steps) required parsimoniously by the considered tree, and m represents the minimum amount of change that the character may show on any tree. Both s and m depend on the suite of terminals used, for which reason it is assumed throughout that some particular set of terminals is treated.

The change, s, in a character on a tree can be partitioned into observed variation,[1] m, and homoplasy (extra steps), h:

$$h = s - m.$$

The consistency index expresses that partition as fractions of the amount of change, s; $(1 - c)$ is the fraction of change that must be attributed to homoplasy. The character fits the tree poorly to the degree that the tree requires homoplasy in the character. When no homoplasy is required, the fit is perfect, and then $c = 1$. As m cannot exceed s, c cannot exceed unity. If $s = 0$, then $m = 0$, and c is taken to be unity.

The amount of homoplasy might also be expressed as a fraction of possible homoplasy. This is done by the distortion coefficient, d, of Farris (1973):

$$d = h/(g - m)$$

where g denotes the greatest amount of change that the character may require on any tree, that is, the greatest possible value of s, so that $(g - m)$ is the greatest possible value of the amount of homoplasy, h. The complement of d, here denoted r, is

$$r = 1 - d = \frac{(g - m) - (s - m)}{g - m} = \frac{g - s}{g - m}.$$

If $g = m$, then $s = g$, and r is taken to be unity, so that d is taken to be 0. I shall call r the *retention index*.

The interpretation of the retention index can be seen from a simple argument. On a tree for which $s = m$, $r = 1$, there is no homoplasy, and all similarities between terminals in the character are homologous. On another tree for which $s > m$, some of those similarities are homoplasies. Each additional requirement for a step implies a separate

---

[1] The observed variation has frequently been called the range of a character. That usage is appropriate for numerical characters, but not for others, such as sequence sites or those having multifurcating character state trees.

0748-3007/89/040417+03 $03.00/0

origin for a state, and each such new origin reduces the fraction of similarities that can be regarded as homologous. If $s=g$, $r=0$, and the character shows as much homoplasy as possible. Similarity in this character is then irrelevant to the groupings of the tree; all apparent apomorphic similarities in the character are dismissed as nonhomologies. The retention index is then the fraction of apparent synapomorphy in the character that is retained as synapomorphy on the tree.

It might be thought that the amount of possible synapomorphy should be measured as g, not as $g-m$, so vitiating that interpretation. But if $g=m$, no homoplasy is possible. All similarities in the character in that case are symplesiomorphies, and any nonplesiomorphic states present are necessarily autapomorphies, so that there is no possible synapomorphy at all. To put this another way, an apparent synapomorphy can be regarded as providing evidence of grouping only if that similarity would have to be attributed to homoplasy on accepting some alternative grouping. There is then no potential for evidence, for synapomorphy, unless there is some potential for homoplasy.

When interpreted as the fraction of total change, s, attributed to observed variation (c) or else to homoplasy ($1-c$), the consistency index seems unexceptionable. But for assessing the fit of a character to a tree, it might be desirable to use a measure that reaches 0 when a character fits a tree as poorly as possible. The consistency index lacks this property, as it can be no less than m/g. An index with the intended behavior can be obtained by a linear rescaling of the consistency. As $(c-m/g)$ varies precisely between $(1-m/g)$ and 0, dividing the first difference by its maximum produces a quantity that varies precisely between 1 and 0.

$$\frac{m/s-m/g}{1-m/g}=\frac{g-s}{g-m}\cdot\frac{m}{s}=rc.$$

I shall call rc the *rescaled consistency index.*

The rescaled consistency can be useful in comparisons of fit of characters with different values of m/g. As an example, observe that a binary character showing the apomorphic state in k terminals, and the plesiomorphic state in k or more, can require at most k steps on any tree, so that $g=k$. Suppose that $s=2$ for each of two binary characters, that $g=2$ for the first, and that $g=11$ for the second. Both have $c=0.5$, and the two are indeed alike in the fraction of change attributable to homoplasy. But the first fits the tree as poorly as possible, while the second could do considerably worse. This is reflected in the rescaled consistencies. For the first character, $rc=0$; for the second, $rc=0.45$.

Successive weighting as implemented in PHYSYS (see Carpenter, 1988; cf. Farris, 1969) used consistencies as weights, but that method can perhaps be improved. It seems desirable to use a weighting function that can reach 0 when the character has as much homoplasy as possible. For this reason I designed Hennig86 to calculate weights from rescaled consistencies.

To compute any of these indices for a suite of characters rather than a single character, each of the quantities m, g, and s is first summed over characters in the suite, yielding totals M, G, S. The ensemble[2] indices are then calculated analogously to those for single characters.

$$C=M/S \qquad R=(G-S)/(G-M)$$

---

[2] The ensemble consistency has also been referred to as the "overall" consistency. This would in general be misleading, as the suite need not comprise all characters—certainly not if "all" is taken literally. That adjective, moreover, has become so closely associated with the pheneticists' notion of "overall" (i.e. raw) similarity, that it is best to avoid the term unless the negative connotation is intended.

The ensemble rescaled consistency is simply *RC*. Because *R* depends only on the totals, uninformative characters, those having $g = m$, do not influence the ensemble retention index. *C* is not affected by characters having $g = 0$, but other characters with $g = m$ push the ensemble consistency closer to unity.

The effect of uninformative characters on *C* is of little importance when different trees are compared on the fit of the same characters, but it may produce misleading results when different suites of characters, or different weightings of the same characters, are compared.

For example, Colless (1980) advocated weighting characters by $1/m$, so that after weighting $m = 1$ for each character.[3] For the morphometric data of Mickevich and Johnson (1976) he reported that $C = 0.66$ without weighting and $C = 0.7$ with it, concluding that his weighting had reduced the fraction of change due to homoplasy. Mickevich and Farris (1981), however, pointed out that two of those six characters were uninformative, and that those characters had smaller *m* than some others, so that the weighting had magnified their influence on *C*. With those two characters deleted they found $C = 0.64$ without the weighting and $C = 0.61$ with it. Colless' scheme had in fact increased the weight of autapomorphies and homoplasies, rather than of synapomorphies, but his comparisons of *C* values seemed superficially to suggest the opposite.

Autapomorphies will not always be so conveniently concentrated in a few characters. In any case Mickevich's and Farris' selection did not entirely remove the effects of autapomorphies on *C*, as unique derivations occurred in other characters as well. For comparisons such as this it would be better to use a measure insensitive to uninformative variation. Using Hennig86, I find $R = 0.80$ without Colless' weighting and $R = 0.76$ with it. The same values are obtained, whether the two uninformative characters are deactivated or not.

Acknowledgment

I wish to thank Dr N. I. Platnick for his extremely useful suggestions.

REFERENCES

CARPENTER, J. M. 1988. Choosing among multiple equally parsimonious cladograms. Cladistics 4: 291–296.

COLLESS, D. H. 1980. Congruence between morphometric and allozyme data for *Menidia* species: a reappraisal. Syst. Zool. 29: 288–299.

FARRIS, J. S. 1969. A successive approximations approach to character weighting. Syst. Zool. 18: 374–385.

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

FARRIS, J. S. In press. Phenetics in camouflage. Cladistics.

KLUGE, A. G. AND J. S. FARRIS. 1969. Quantitative phyletics and the evolution of anurans. Syst. Zool. 18: 1–32.

MICKEVICH, M. F. AND J. S. FARRIS. 1981. The implications of congruence in *Menidia*. Syst. Zool. 30: 351–370.

MICKEVICH, M. F. AND M. S. JOHNSON. 1976. Congruence between morphological and allozyme data in evolutionary inference and character evolution. Syst. Zool. 25: 260–270.

Note Added in Proof

Over a year after the release of Hennig86, Archie (1989, Syst. Zool. 38: 253–269) proposed the "homoplasy excess ratio maximum", which is identical to the ensemble retention index. That cumbersome terminology is rather ill-chosen, since the index attains its maximum when there is no homoplasy.

---

[3] Such weighting has been recommended by other pheneticists as well. As one might expect, it does not turn out to be a particularly sensible idea. For discussions of this point see Mickevich and Farris (1981) and Farris (in press).