FORUM

Cladistics (1995) 10:315–319

TESTING SIGNIFICANCE OF INCONGRUENCE

James S. Farris,[1,2] Mari Källersjö,[1] Arnold G. Kluge[3] and Carol Bult[4]

[1] Naturhistoriska riksmuseet, Molekylärystematiska laboratoriet, Box 50007, S-104 05, Stockholm, Sweden; [2] Department of Entomology, American Museum of Natural History, Central Park West at 79th St., New York, New York 10024, U.S.A.; [3] Department of Reptiles and Amphibians, Museum of Zoology, The University of Michigan, Ann Arbor, Michigan 48109, U.S.A. and [4] Institute of Genomic Research, 932 Clopper Rd., Gaithersburg, Maryland 20878, U.S.A.

Received for publication 22 December 1994; accepted 19 January 1995

Introduction

Assessments of incongruence between sources of phylogenetic evidence have often been largely intuitive, based simply on inspecting results from different data matrices. Even conclusions based on quantitative indices have generally rested on essentially arbitrary interpretation; there have been no criteria for deciding how large an index would be needed to demonstrate incongruence convincingly. Here we show how to overcome this difficulty by constructing a statistical test of the null hypothesis of congruence[1].

Partitions

The problem can be divided into two parts, selecting a measure of incongruence, and obtaining the null distribution of the measure. We discuss choice of a measure in the next section. For the distribution we use random partitioning, which could be applied to any of several measures.

The principle is conveniently illustrated by the familiar Mann-Whitney *U* test (cf. Lindgren, 1962; 352) of the null hypothesis that two populations do not differ in location. The *U* statistic is just the number of comparisons in which an observation from the first sample exceeds one from the second, and for a one-tailed test the null hypothesis can be rejected when *U* is large enough. Just how large is large enough is determined from the null distribution; the critical value is picked so that the chance of rejecting incorrectly is no greater than the desired significance level, usually 5%.

The null distribution is obtained by regarding the two actual samples with numbers of observations M and N as one possible partition of the total of $(M + N)$ observa-

[1] JSF presented this test first at the 1991 meeting of this Society and later at the 1993 Nordic Phylogenetic Systematics Network meeting. CB presented papers using this test at the 1991 WHS meeting, at the Smithsonian Institution's Laboratory of Molecular Systematics and at George Washington University. The procedure was also described with the permission of the authors by J. M. Carpenter at the 1991 meeting of the Entomological Society of America, by A. Tehler at the Fifth International Mycological Congress, and by A. Bruneau at the 1994 ESF Workshop on Molecular and Classical Taxonomy. At the 1994 ESA meeting, C. W. Cunningham described our procedure as having been developed by Swofford from "ideas in Schuh and Farris (1981)". Apparently those ideas occurred to Swofford some time after he attended CB's presentation of our method at the Smithsonian in 1992.

0748-3007/94/030315+05$08.00/0

tions. Every possible way of assigning the available observations to groups of the same two sizes is considered equally likely under the null hypothesis. The probability of a certain U value under the null hypothesis, then, is found by counting the number of such partitions which would yield that U.

Conceptually, the same approach is easy to apply to testing for incongruence between sets of characters (or sites). The null hypothesis of congruence would be rejected if the measure D of incongruence is large enough. Data matrices with M and N characters for the same taxa can be regarded for this purpose as a possible partition of $(M+N)$ total characters. The null distribution of D would be determined by summing over possible partitions of those characters into suites of sizes M and N.

In practice, however, incongruence is much more laborious to calculate than is U, while M and N may easily be large enough so that the number of possible partitions[2] is too great for direct enumeration. Fortunately, to obtain a significance test it is only necessary to compute D for a relatively small number of partitions, these being chosen at random from among those possible.

To perform the test, the value of D is found for the observed data matrices, then for a number W of randomly-selected partitions of the characters into matrices of the two original sizes. If a number S of the D values from randomly-selected partitions are smaller than the observed D, then the Type I error rate (tail probability) on rejecting the null hypothesis is $1 - S/(W+1)$. If $W=99$ and $S=95$, for example, this indicates significance at the 5% level.

As with U, the degree of departure between two available samples—sets of characters—is judged convincing if the observed difference—incongruence—is larger than would be likely to be encountered among other partitions of the data.

Length Differences

Early attempts to quantify congruence were based on trees as such. Trees obtained from different matrices might be compared, for example, by counting the groups on their consensus. Many indices of that type have been proposed, but all such approaches share the serious weakness that they take no account of strength of evidence. A group makes the same contribution to the index, whether it is supported by many characters or only a few.

That is unsuitable for purposes of assessing incongruence between suites of characters. A group strongly supported by one matrix and strongly disputed by another reflects equally strong disagreement between those matrices. But no substantial disagreement is indicated by a group that is only weakly supported (or disputed) by one of the matrices.

Mickevich and Farris (1981), who called attention to this problem[3], introduce a measure that avoids that difficulty. For matrices X and Y the incongruence length difference[4] D_{XY} is given by:

$$D_{XY} = L_{(X+Y)} - (L_X + L_Y) \tag{1}$$

[2] When $M=25$, $N=26$, there are $(M+N)!/M!N! = 2.5 \times 10^{14}$.

[3] Swofford (1991: 311), who also discussed the Mickevich/Farris measure, nonetheless attributed this criticism of consensus methods to Miyamoto (1985). Miyamoto (1985: 186) cited Mickevich and Farris (1981) and Schuh and Farris (1981).

[4] The name is new; Mickevich and Farris provided none.

L_X, L_Y and $L_{(X+Y)}$ denote the lengths of most parsimonious trees calculated for each matrix separately and for the combined matrix, that including all the characters.

D_{XY} is 0 when there is at least one tree on which the two matrices agree. It is large when minimizing homoplasy in one set of characters can be accomplished only by increasing homoplasy in the other set. This occurs to the extent that groups well supported (in the sense of Bremer's (1988) length difference)[5] by one matrix conflict with groups well supported by the other.

D_{XY} is particularly suitable for measuring incongruence between matrices because it isolates that source of incongruence. Incongruence among the characters of either original matrix is included as homoplasy in one of the single-matrix lengths L_X and L_Y. The combined matrix length $L_{(X+Y)}$ exceeds $(L_X + L_Y)$ only to the degree that further homoplasy arises when the original matrices are brought together.

A simplification of calculations is possible when applying D_{XY} in the random-partitioning test described above. Say that original matrices X, Y are randomly repartitioned into matrices P, Q of the original sizes. The count S used in the test is the number of such repartitions for which:

$$[L_{(X+Y)} - (L_X + L_Y)] > [L_{(P+Q)} - (L_P + L_Q)] \tag{2}$$

But $L_{(P+Q)} = L_{(X+Y)}$ because there is just one matrix comprising all the characters. On eliminating the common term, S is found to be the number of repartitions for which:

$$(L_X + L_Y) < (L_P + L_Q) \tag{3}$$

To perform the test, then, $L_{(X+Y)}$ need not be evaluated at all.

The random partitioning technique is not limted to two samples, and the simplicity of (3) makes it easy to extend to a test of congruence among several matrices. For three original matrices X, Y, Z, S would be the number of random repartitions into matrices P, Q, R of the original sizes for which:

$$(L_X + L_Y + L_Z) < (L_P + L_Q + L_R) \tag{4}$$

As before, reject when the error rate $1 - S/(W+1)$ is small enough.

The lengths used in the test could be found from exact most-parsimonious trees for each matrix, but that is unnecessarily time-consuming. As the test uses just the number S of repartitions satisfying inequality (3) and (4), exact lengths are not crucial; it is enough to get the sums in the right order. It should then generally be adequate to use fast, approximate parsimony calculations, allowing a great reduction in effort.

Applications

We have incorporated these ideas into a program, arn[6], which reads a combined matrix and a description of the actual partition of characters, then generates random repartitions and carries out the congruence test automatically. Both two-sample and multi-sample tests are provided.

---

[5] Now called Bremer support (see Källersjö et al., 1992; Farris et al., 1994).

[6] For Arnold. Described by JSF at the 1991 meeting of this Society; demonstrated by CB at the Smithsonian Institution's Laboratory of Molecular Systematics. A second program, kon, which performs the same test but has a different output format, was demonstrated by JSF and used in a workshop at the 1993 meeting of the Nordic Phylogenetic Systematics Network. K. Nixon has, with the permission of the authors, also included our test in his program DADA.

This is a prototype, and its efficiency can be substantially improved, but it is already fast enough to make the congruence test easy to use. The allozyme and morphological data compared by Mickevich and Farris (1981) have a total of 68 characters for 16 taxa. For those data, it takes arn 30 seconds on a 66 MHz 80486DX2 to perform the congruence test with $W = 999$ random repartitions. For Kluge's (1989) morphological and biochemical data, with a total of 77 characters for 11 taxa, a similar analysis requires 20 seconds.

This program has been applied in several studies by other authors. Bremer and Struwe (1992) used it to contrast morphological and chemical characters. Bruneau and Dickson (1995) performed a three-matrix congruence test with isozyme, morphological, and restriction-site data. Smith and Sytsma (1994) compared morphological and nucleotide-sequence data, as did Tehler (1994, 1995a, b).

Discussion

Swofford (1991: 316) was unaware of our test procedure, but his objections to the Mickevich/Farris measure might appear relevant.

> "One difficulty is the need to combine the data sets into a single matrix. If, for example, one data set were cranial morphology and the other ribosomal RNA (rRNA) sequences, the morphological characters could easily be overwhelmed by the large number of molecular characters".

In fact, however, he has confused two quite distinct issues. That a large matrix could overwhelm a small one might arise as a concern when inferring a tree from the combined matrix[7], but it is hardly a problem when assessing incongruence between matrices.

Swofford's misgiving presupposes that the two kinds of characters are incongruent. If they were congruent, their relative numbers would make no difference—there would be no opposition to overwhelm. But if they are incongruent, abundance of one of the kinds scarcely hinders recognition of that fact.

Suppose for example that matrix X comprises two binary characters, each splitting the four terminals AD/BC, while matrix Y includes 50 binary characters, each of which splits the terminals AB/CD. Overwhelming though this seems, our test readily finds the incongruence between X and Y to be very highly significant, the tail probability being 0·00075[8].

Swofford (1991: 316) used an example of his own to illustrate "a potentially more serious limitation". His X has 10 AB/CD characters, while his Y has five AB/CDs and five AD/BCs. D_{XY} is then 0, which he protested (p. 317):

> "We would therefore conclude that there is no between-data-set incongruence, even though half the characters in data set Y contradict the tree unanimously supported by the characters in data set X. This seems unreasonable."

This time he has confused between- and within-matrix incongruence, and this can be seen in two ways. First, some of the characters in Y dispute those in X, but then they also dispute the rest of Y in just the same way. Their disagreement "with X" is just the incongruence within Y stated again. All the homoplasy in the entire analysis is already present when Y is analyzed alone.

---

[7] Though even this is doubtful.

[8] This probability was not found by randomly sampling partitions, but from the exact distribution of D_{XY}, the latter calculation being feasible in this simple case.

Second, considered in isolation, some of the characters in *Y* dispute those in *X*, but then just as many others agree with those in *X*, and the net effect is 0. Neither faction by itself constitutes the relationship between *matrix Y* and *X*. *Matrix Y* does not dispute *X* at all, for the simple reason that *Y* is—collectively—neutral with respect to *X*.

This all suggests a last analogy with the Mann-Whitney *U*, which is intended to evaluate variation between samples. Given two samples with the same mean, *A* with zero variance, *B* with great variance, *U* will quite correctly fail to find a difference. "But this seems unreasonable"—someone might say—"Most of the observations in *B* differ greatly from the value unanimously supported by the observations in *A*."

Testing for incongruence between matrices involves measuring just that, and our approach appears to provide an effective means of doing so.

Acknowledgments

This work was supported in part by NFR grants 102044-300 to J. S. Farris and 09858-303 to M. Källersjö. We thank Dr D. Eernisse for his strong and timely encouragement.

REFERENCES

BREMER, K. 1988. The limits of amino-acid sequence data in angiosperm phylogenetic reconstruction. Evolution 42: 795–803.

BREMER, B. AND L. STRUWE. 1992. Phylogeny of the Rubiaceae and the Loganiaceae: congruence or conflict between morphological and molecular data? Amer. Jour. Bot. 79: 1171–1184.

BRUNEAU, A. AND E. E. DICKSON. 1995. Congruence of chloroplast DNA restriction site characters with morphological and isozyme data in *Solanum* sect. *Lasiocarpa*. Can. Jour. Bot. (in press).

FARRIS, J. S., M. KÄLLERSJÖ, A. G. KLUGE AND C. BULT. 1994. Permutations. Cladistics 10: 65–76.

KLUGE, A. G. 1989. A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). Syst. Zool. 38: 7–25.

KÄLLERSJÖ, M., J. S. FARRIS, A. G. KLUGE AND C. BULT. 1992. Skewness and permutation. Cladistics 8: 275–287.

LINDGREN, B. W. 1962. Statistical Theory. MacMillan, New York.

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

MIYAMOTO, M. 1985. Consensus cladograms and general classifications. Cladistics 1: 186–189.

SCHUH, R. T. AND J. S. FARRIS. 1981. Methods for investigating taxonomic congruence and their application to the Leptopodomorpha. Syst. Zool. 30: 331–351.

SMITH, J. F. AND K. J. SYTSMA. 1994. Molecules and morphology: congruence of data in *Columnea* (Gesneriaceae). Plant Syst. Evol. (in press).

SWOFFORD, D. L. 1991. When are phylogeny estimates from molecular and morphological data incongruent? *In* M. M. Miyamoto and J. Cracraft (eds), Phylogenetic Analysis of DNA Sequences. Oxford Univ. Press, New York, pp. 295–333.

TEHLER, A. 1994. Cladistic analysis in ascomycete systematics: Theory and practice. *In* D. L. Hawksworth (ed.), Ascomycete Systematics: Problems and Perspectives in the 90s. Plenum, New York, pp. 185–197.

TEHLER, A. 1995a. Arthionales phylogeny as indicated by morphological and rDNA sequence data. Cryptogamic Botany (in press).

TEHLER, A. 1995b. Morphological data, molecular data, and total evidence in phylogenetic analysis. Can. Jour. Bot. (in press).