chromosomal, genic, and classical data sets. J. Mammal. 65:643–654.
SULLIVAN, J., K. E. HOLSINGER, AND C. SIMON. 1995. Among-site rate variation and phylogenetic analysis of 12S rRNA in sigmodontine rodents. Mol. Biol. Evol. 12:988–1001.
SWOFFORD, D. L. 1991. When are phylogeny estimates from molecular and morphological data incongruent? Pages 295–333 *in* Phylogenetic analysis of DNA sequences (M. M. Miyamoto and J. Cracraft, eds.). Oxford Univ. Press, New York.
SWOFFORD, D. L. 1996. PAUP*: Phylogenetic analysis using parsimony, version 4.0. Sinauer, Sunderland, Massachusetts.
UZZELL, T., AND K. W. CORBIN. 1971. Fitting discrete distributions to evolutionary events. Science 172: 1089–1096.
YANG, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. J. Mol. Evol. 39: 306–314.
YANG, Z., N. GOLDMAN, AND A. FRIDAY. 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. Mol. Biol. Evol. 11:316–324.
YANG, Z., I. J. LAUDER, AND H. J. LIN. 1995. Molecular evolution of the hepatitus B virus genome. J. Mol. Evol. 41:587–596.

*Received 12 September 1995; accepted 5 March 1996*
*Associate Editor: David Cannatella*

*Syst. Biol.* 45(3):380–385, 1996

# The Probabilistic Basis of Jaccard's Index of Similarity

RAIMUNDO REAL AND JUAN M. VARGAS

*Department of Animal Biology, Faculty of Science, University of Málaga, Málaga 29071, Spain;*
*E-mail: rrgimenez@ccuma.uma.es (R.R.)*

Interspecific association analysis from presence/absence data is an unresolved topic in ecology and biogeography (e.g., Connor and Simberloff, 1979, 1983, 1984, 1986; Simberloff and Connor, 1981; Gilpin and Diamond, 1982, 1984; Ryti and Gilpin, 1987; Jackson et al., 1992). Several techniques have been proposed to test association between species. Connor and Simberloff (1979) put forward a null model based on the Monte Carlo randomization procedure. Gilpin and Diamond (1982) took a different approach, based on a log-linear model with binary data. Jackson et al. (1992) proposed a hybrid model combining the two previous methods. However, all of these null models use an observed data matrix to generate a null distribution, and so observed and null distributions lack statistical independence (Grant and Abbott, 1980). In a fourth approach, now called the coefficient model (see Jackson et al., 1992), the observed distribution of a similarity index is tested against a distribution of expected values for that index.

In this context, similarity indices are frequently used to study the coexistence of species or the similarity of sampling sites. A matrix of similarity coefficients, between either species or locations, may be analyzed in two ways: by ordination, i.e., by attempting to arrange the locations or species within a theoretically continuous sequence, or by classification, the aim of which is to place the locations or species in discontinuous groups (McCoy et al., 1986), which may overlap in nonhierarchical classification approaches. The main aim of this type of analysis is to discover distribution patterns common to different species and groups of areas with similar biota (Birks, 1987). However, Simberloff and Connor (1979) stated that most indices of similarity are not associated with probability values because their underlying distributions are unknown, thus preventing high and low levels of association between species from being recognized objectively with regard to what may be expected at random. Only the distributions of the simple matching coefficient (Goodall, 1967), the Baroni-Urbani and Buser coefficient (Baroni-Urbani and Buser, 1976), the Jaccard coefficient (Baroni-Urbani, 1980), and the phi coefficient (Jackson et al., 1992) have been discussed in the literature.

Jaccard's index (Jaccard, 1908) stands out as one of the most useful and widely used indices of the 60 or so similarity indices for binary data (Birks, 1987). Moreover, it can be used in species conservation because it may be applied to the power function of the relationship between species and areas to determine a measure for the optimum size for natural protection reserves (Higgs and Usher, 1980). Jaccard's index does not take into account negative matches. In this way the similarity between two operational taxonomic units (OTUs) is not influenced by other OTUs included in the analysis, and the value of Jaccard's index is independent of the number of OTUs studied. However, those indices taking negative matches into account change their values when additional OTUs, with species absent from the two OTUs analyzed, are included in the analysis (Buser and Baroni-Urbani, 1982).

As Rice and Belland (1982) pointed out, it would not be correct to infer close biological similarity directly from high values of Jaccard's index nor to infer biological dissimilarity from low values, because these could be random. In turn, the random values expected to occur will depend on the number of attributes present in the sets formed by each pair of OTUs. Therefore, it is necessary to determine whether the values of Jaccard's index in each pair of OTUs compared differ from what would be expected at random in order to infer their biological significance.

In this paper, we reexamine the probability table associated with Jaccard's similarity index as provided by Baroni-Urbani (1980) and analyze the probabilities associated with this index according to (1) the total number of attributes in both OTUs and (2) the number of attributes in each OTU. We also explore the implications for the use of this index in biogeographical studies.

## Probability Basis of Jaccard's Index

Jaccard's index may be expressed in several ways. A common approach is the following:

$$J = \frac{C}{A + B - C}, \qquad (1)$$

in which $A$ is the number of attributes present in OTU a, $B$ is the number of attributes present in OTU b, and $C$ is the number of attributes present in both OTUs a and b.

Jaccard's index can also be expressed thus:

$$J = \frac{C}{A + B + C}, \qquad (2)$$

where $A$ is the number of attributes present in OTU a and absent in OTU b, $B$ is the number of attributes present in OTU b and absent in OTU a, and $C$ is the same as in Equation 1.

A third way of expressing Jaccard's index is as follows:

$$J = \frac{C}{N}, \qquad (3)$$

in which $C$ is the same as in Equations 1 and 2 and $N$ is the total number of attributes found in both OTUs together.

Jaccard's index demands that $A$, $B$, and $C$ take values from the set of natural numbers, which leads to a distribution in $J$ that is not continuous. Therefore, to calculate values higher and lower than those expected at random, it is necessary to use probability calculus instead of analyzing a continuous statistical distribution of the index.

Unlike other indices, Jaccard's was studied by Baroni-Urbani (1980) from a statistical point of view, and he obtained a statistical table of associated probabilities. Baroni-Urbani (1980) calculated the statistical table for Jaccard's index starting from the probabilities associated with a specific combination ($A$, $B$, $C$, $D$) in the Baroni-Urbani and Buser index:

$$B = \frac{\sqrt{C*D} + C}{\sqrt{C*D} + A + B + C}, \qquad (4)$$

where *A*, *B*, and *C* are the same as in Jaccard's index (Eq. 2) and *D* represents double absences.

In effect, the likelihood that a specific combination (*A*, *B*, *C*, *D*) appears on the Baroni-Urbani and Buser index is

$$P(A, B, C, D) = \frac{N!}{A!B!C!D!} \times 2^{-2N} \quad (5)$$

(see Baroni-Urbani and Buser, 1976). The first factor in this formula corresponds to cases in which the *A*, *B*, *C*, *D* combination is favored and is derived from the permutations with repetition of *N* elements, within which *A*, *B*, *C*, and *D* elements are equal, resulting in the following:

$$P_N^{A,B,C,D} = \frac{N!}{A!B!C!D}. \quad (6)$$

The second factor corresponds to all possible cases that result from the variations with repetitions of the four elements (*A*, *B*, *C*, *D*) taken in groups of *N* elements:

$$VR4, N = 4^N = 2^{2N}, \quad (7)$$

which is why the likelihood of finding an *A*, *B*, *C*, *D* combination in the Baroni-Urbani and Buser index is

$$P(A, B, C, D) = \frac{P_N^{A,B,C,D}}{VR4, N} = \frac{\frac{N!}{A!B!C!D!}}{2^{2N}} = \frac{N!}{A!B!C!D!} \times 2^{-2N}. \quad (8)$$

From this point on, Baroni-Urbani (1980) considered that the probabilities associated with Jaccard's index are similar to these, but with *D* = 0. However, he only took this difference into account when he calculated *N* = *A* + *B* + *C* + *D* in the following way:

$$P(A, B, C, D) = \frac{(A + B + C + D)!}{A!B!C!D!} \times 2^{-2(A+B+C+D)}, \quad (9)$$

which is why, on making *D* = 0, he assumed that

$$\begin{aligned} P(A, B, C, D = 0) &= \frac{(A + B + C + 0)!}{A!B!C!1} \times 2^{-2(A+B+C+0)} \\ &= \frac{(A + B + C)!}{A!B!C!} \times 2^{-2(A+B+C)} \\ &= \frac{N!}{A!B!C!} \times 2^{-2N}. \end{aligned} \quad (10)$$

However, by making *D* = 0, the number of elements to combine is also reduced, from four in the Baroni-Urbani and Buser index to three in Jaccard's index (*A*, *B*, *C*). This is why the factor that calculates the number of possible cases

$$2^{-2N} = 4^{-N} \quad (11)$$

in the previous equations, where the number 4 represents the four elements to be combined in the Baroni-Urbani and Buser index, must change to

$$3^{-N}. \quad (12)$$

Here the number 3 refers to the three elements (*A*, *B*, *C*) combined in Jaccard's index. For this reason, the probability associated with an *A*, *B*, *C* combination in Jaccard's index is the result of dividing the number of cases in which this combination is favored

$$P_N^{A,B,C} = \frac{N!}{A!B!C!} \quad (13)$$

by the number of possible cases

$$VR3, N = 3^N \quad (14)$$

so that the corrected formula is

$$P(A, B, C) = \frac{N!}{A!B!C!} \times 3^{-N}, \quad (15)$$

which must then substitute for Equation 10.

Departing from this point, when the accumulated probabilities are calculated, the probability of finding a random value for *J* that is less than or equal to a given *J*, as expressed in Equation 2, is as follows:

$$P = \frac{\sum_{x=0}^{C} \binom{N}{x} VR2, N - x}{VR3, N}, \quad (16)$$

and the probability of finding a random value for $J$ equal to or greater than a given $J$ is

$$P = 1 - \frac{\sum_{x=0}^{C-1} \binom{N}{x} VR2, N - x}{VR3, N}. \quad (17)$$

Calculations of these formulas can be tabulated for different values of $N$ if frequent use of the probabilities is required.

When these probabilities are calculated in this way, the possibility of finding equally each attribute only in OTU a, only in OTU b, or in both OTUs is taken into account. Thus, in extreme cases all the attributes may be found either exclusively in OTU a or exclusively in OTU b. In other words, the attributes can change indiscriminately from absence to presence or from presence to absence. This is the usual approach in the coefficient model, where marginal totals, in this case the number of attributes in OTU a and OTU b, are not maintained. In this way, the probability $P$ for each species being present in a region is 1/2; $C$ has a binomial distribution, and $J$, as expressed in Equation 3, has the distribution of a binomial random variable divided by the number of trials, in this case $N$, when the probability of a species being present in both OTUs is 1/3. Without any other additional knowledge, one could assume that species not shared by the OTUs should have equal probability of occurring in either OTU, rather than restricting them to a particular OTU, and then, if the test is based on random association between species across the OTUs, one would expect equal probabilities for site occupancy. Thus, with these assumptions, Equations 16 and 17 would be equivalent to one-tailed binomial tests, which is useful because a one-tailed test is considered appropriate for testing ecological and biogeographical hypotheses (see Grant and Abbot, 1980).

However, in certain biogeographical analyses it may not make sense to assume that species not shared by the two OTUs can be equally found in one OTU or the other but rather that they are more likely to be found in the OTU that actually supports more species. Therefore, it would not be reasonable to take this possibility into account in the calculation of probabilities. Jackson et al. (1992), for instance, considered the coefficient model to be less conservative than other models because marginal totals are not maintained. To maintain marginal totals, it may be assumed that a species present in OTU a could enter OTU b and therefore be present in both OTUs but will not disappear from OTU a. In this case, the attribute would change from absent to present, but not vice versa. The problem is analogous to that of the different interpretations given of shared species in cladistic biogeography (see Page, 1988). Thus, the probability formulas developed so far would take into account a situation analogous to that considered in the Wagner parsimony criterion (Kluge and Farris, 1969; Farris, 1970), which deals with binary conditions and allows free reversibility. However, the biogeographer may consider the species distribution between OTUs similar to that considered in the Camin–Sokal parsimony criterion (Camin and Sokal, 1965), which does not allow reversals from a derived state, such as presence, to a more ancestral state, such as absence.

On the assumption that the conditions are irreversible, it would be more appropriate to calculate the probabilities associated with Jaccard's index by fixing the total number of elements, $A$ and $B$, in each OTU. Strauss (1982) considered fixing the number of species in each location a realistic approach applicable to any similarity index (see also Connor and Simberloff, 1979, 1984; Diamond and Gilpin, 1982; Gilpin and Diamond, 1982). This approach is even more pertinent in the case of Jaccard's index because it is influenced by the size of the sample and tends to group together the OTUs that have similar elements and a similar number of species (Sepkoski and Rex, 1974; Connor and Simberloff, 1978).

If presences are considered irreversible, the number of possible cases would be the sum of the groups of common elements that could be formed while preserving the number of elements $A$ and $B$ of each OTU, with $J$ expressed as in Equation 1:

$$\sum_{x=0}^{\mathrm{Min}(A,B)} \binom{A + B - x}{x}. \qquad (18)$$

Thus, the common elements may range from 0 to the minimum value $A$ and $B$. The cases in which a certain $A$, $B$, $C$ combination is favored would result from the combinations of $C$ elements that can be formed from the total of the $A + B - C$ elements:

$$C^{C}_{A+B+C} = \binom{A + B - C}{C}. \qquad (19)$$

When the accumulated probabilites are calculated, the likelihood of finding a random value of $J$, as expressed in Equation 1, greater than or equal to a given $J$ value is as follows:

$$P = 1 - \frac{\displaystyle\sum_{x=0}^{C-1} \binom{A + B - x}{x}}{\displaystyle\sum_{x=0}^{\mathrm{Min}(A,B)} \binom{A + B - x}{x}}. \qquad (20)$$

The likelihood of randomly finding a $J$ value less than or equal to a given $J$ value is as follows:

$$P = \frac{\displaystyle\sum_{x=0}^{C} \binom{A + B - x}{x}}{\displaystyle\sum_{x=0}^{\mathrm{Min}(A,B)} \binom{A + B - x}{x}}. \qquad (21)$$

These equations can be used to calculate the significance of Jaccard's similarity coefficients while maintaining a fixed total number of attributes, $A$ and $B$, in each of the OTUs being compared. The result of this operation can be tabulated, although for a given total number of attributes, $N$, the probabilities vary according to the number of attributes of the OTU with the fewest attributes.

Hence, either Equations 16 and 17 or Equations 20 and 21 can be used to calculate the probabilities associated with a value in Jaccard's index. If the probability is calculated by fixing a set number of total attributes in each OTU (irreversible conditions), Equations 20 and 21 should be used. Alternatively, if any possible distribution for the $N$ elements in both OTUs is considered (reversible conditions), Equations 16 and 17 should be used.

## Acknowledgments

We thank Cesare Baroni-Urbani and Agustín Antúnez for their improvements on an earlier version of the manuscript. This work was partially supported by DGICYT, project PB 93-1002.

## References

Baroni-Urbani, C. 1980. A statistical table for the degree of coexistence between two species. Oecologia 44:287–289.

Baroni-Urbani, C., and M. W. Buser. 1976. Similarity of binary data. Syst. Zool. 25:251–259.

Birks, H. J. B. 1987. Recent methodological developments in quantitative descriptive biogeography. Ann. Zool. Fenn. 24:165–178.

Buser, M. W., and C. Baroni-Urbani. 1982. A direct nondimensional clustering method for binary data. Biometrics 38:351–360.

Camin, J. H., and R. R. Sokal. 1965. A method for deducing branching sequences in phylogeny. Evolution 19:311–326.

Connor, E. F., and D. Simberloff. 1978. Species number and compositional similarity of the Galapagos flora and avifauna. Ecol. Monogr. 48:219–248.

Connor, E. F., and D. Simberloff. 1979. The assembly of species communities: Chance or competition? Ecology 60:1132–1140.

Connor, E. F., and D. Simberloff. 1983. Intraspecific competition and species co-occurrence patterns on islands: Null models and the evaluation of evidence. Oikos 41:455–465.

Connor, E. F., and D. Simberloff. 1984. Neutral models of species' co-occurrence patterns. Pages 316–331 *in* Ecological communities: Conceptual issues and the evidence (D. R. Strong, Jr., D. Simberloff, L. G. Abele, and A. B. Thistle, eds.). Princeton Univ. Press, Princeton, New Jersey.

Connor, E. F., and D. Simberloff. 1986. Competition, scientific method, and null models in ecology. Am. Sci. 74:155–162.

Diamond, J. M., and M. E. Gilpin. 1982. Examination of the "null model" of Connor and Simberloff for species co-occurrences on islands. Oecologia 52:64–74.

Farris, J. S. 1970. Methods for computing Wagner trees. Syst. Zool. 19:83–92.

Gilpin, M. E., and J. M. Diamond. 1982. Factors contributing to non-randomness in species co-occurrences on islands. Oecologia 52:75–84.

Gilpin, M. E., and J. M. Diamond. 1984. Are species co-occurrences on islands non-random, and are null hypotheses useful in community ecology? Pages

297–315 *in* Ecological communities: Conceptual issues and the evidence (D. R. Strong, Jr., D. Simberloff, L. G. Abele, and A. B. Thistle, eds.). Princeton Univ. Press, Princeton, New Jersey.

GOODALL, D. W. 1967. The distribution of the matching coefficient. Biometrics 23:647–656.

GRANT, P. R., AND I. ABBOTT. 1980. Interspecific competition, island biogeography and null hypotheses. Evolution 34:332–341.

HIGGS, A. J., AND M. B. USHER. 1980. Should reserves be large or small? Nature 285:568–569.

JACCARD, P. 1908. Nouvelles recherches sur la distribution florale. Bull. Soc. Vaudoise Sci. Nat. 44:223–270.

JACKSON, D. A., K. M. SOMERS, AND H. H. HARVEY. 1992. Null models and fish communities: Evidence of nonrandom patterns. Am. Nat. 139:930–951.

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

MCCOY, E. D., S. S. BELL, AND K. WALTERS. 1986. Identifying biotic boundaries along environmental gradients. Ecology 67:749–759.

PAGE, R. D. M. 1988. Quantitative cladistic biogeography: Constructing and comparing area cladograms. Syst. Zool. 37:254–270.

RICE, J., AND R. J. BELLAND. 1982. A simulation study of moss floras using Jaccard's coefficient of similarity. J. Biogeogr. 9:411–419.

RYTI, R. T., AND M. E. GILPIN. 1987. The comparative analysis of species occurrence patterns on archipelagos. Oecologia 73:282–287.

SEPKOSKI, J. J., AND M. A. REX. 1974. Distribution of freshwater mussels: Coastal rivers as biogeographic islands. Syst. Zool. 23:165–188.

SIMBERLOFF, D., AND E. F. CONNOR. 1979. Q-mode and R-mode analyses of biogeographic distributions: Null hypotheses based on random colonization. Pages 123–138 *in* Contemporary quantitative ecology and related econometrics (G. P. Patil and M. Rosenzweig, eds.). International Co-operative, Fairland, Maryland.

SIMBERLOFF, D., AND E. F. CONNOR. 1981. Missing species combinations. Am. Nat. 118:215–239.

STRAUSS, R. E. 1982. Statistical significance of species clusters in association analysis. Ecology 63:634–639.

*Received 27 June 1995; accepted 6 March 1996*
*Associate Editor: Richard Olmstead*

*Syst. Biol.* 45(3):385–390, 1996

# A Further Note on Symmetry of Taxonomic Trees

DONALD H. COLLESS

*Division of Entomology, CSIRO, Canberra 2601, Australia; E-mail: donc@ento.csiro.au*

In a recent paper (Colless, 1995), I presented evidence that cladograms constructed by the Wagner method (Farris, 1970) are inherently less symmetrical than phenograms constructed using a space-conserving algorithm (e.g., WPGMA or UPGMA; Colless, 1995). I there used an index of symmetry, $I_{SYM}$, and fully described that index and the notion of space distortion. One especially intriguing feature of those results was that the difference between cladograms and phenograms holds even for random data, where there could be no explanation in "real" underlying taxonomic patterns. It is becoming clear that for the cladistic case at least this difference is maintained because the algorithm concerned is mimicking a particular model of random evolution (Heard and Mooers, 1996). Here, I comment further on this fact and present additional empirical results.

Other authors (e.g., Slowinski, 1990; Rogers, 1994) have advanced two different models that might be used to calculate theoretically the average $I_{SYM}$ for random trees with various numbers of OTUs: (1) the equal-rates Markov (ERM) model, based on a hypothesis of random speciation with rates uniform across lineages, and (2) the equal probability (EP) model, which assumes that all possible labeled trees are equally probable. The EP model clearly applies to trees constructed by parsimony using random data (Slowinsky, 1990; Mooers et al., 1995; Heard and Mooers, 1996). Like-