*Ann. Bot. Fennici* 40: 415–421 ISSN 0003-3847
Helsinki 17 December 2003 

# Estimating the effect of the similarity coefficient and the cluster algorithm on biogeographic classifications

Miguel Murguía[1] & José Luis Villaseñor[2]

[1)] *Asociación de Biólogos Amigos de la Computación (ABACo, A. C.), Av. San Jerónimo 507, Col. San Jerónimo Lídice, 10200 México, D.F. México (e-mail: mmr@fciencias.unam.mx)*
[2)] *Instituto de Biología, Universidad Nacional Autónoma de México, Departamento de Botánica, Apartado postal 70-237, 04510 México, D.F. México (e-mail: vrios@mail.ibiologia.unam.mx)*

*Received 25 Sep. 2002, revised version received 4 June 2003, accepted 13 June 2003*

Murguía, M. & Villasenõr, J. L. 2003: Estimating the effect of the similarity coefficient and the cluster algorithm on biogeographic classifications. — *Ann. Bot. Fennici* 40: 415–421.

Similarity matrices obtained using a null model and nine similarity coefficients based on an exhaustive and hypothetical set of presence/absence data matrices are generated and compared. Likewise, the biogeographic classifications obtained from an empirical set of data (the genera of Asteraceae of Mexico) and from the application of the same nine similarity coefficients and three cluster methods are compared. It is concluded that differences in the classifications generated from different similarity coefficients can be of almost 50% with the set of hypothetical matrices and more than 70% with the empirical data. The kind of clustering method (single, complete, or average) also generates differences in the classification topologies, even when using the same similarity coefficient. The empirical data produced similar topologies in 51% to 85% of the cases. Due to the dependence among the similarity coefficients, the cluster method used, and the generated classifications, it is concluded that classifications obtained through the use of different similarity coefficients or cluster methods are not comparable. The most similar classification topologies were obtained from the use of the Jaccard and Sorensen-Dice similarity coefficients. They also showed the lowest number of poorly informative structures.

Key words: Asteraceae, cluster analyses, null models, quantitative biogeography, similarity coefficients

## Introduction

Each day more precise biogeographic data are available. They have also been ever more frequently used to classify areas based on knowledge of the distribution patterns of their biota.

Numerous tools are currently available for applying numerical methods to biogeographic analyses. They have contributed to background knowledge that better describes and explains the geographical distribution patterns of organisms. However, analyses carried out with different numerical methods do not always arrive at the same result, mostly due to the variables selected (Crovello 1981, Hubálek 1982). Therefore, understanding the behavior of a particular method using a given set of variables is a crucial issue.

Crovello (1981) pointed out that quantitative biogeographic analysis is a decision making process. Among the decisions to be made are the definition of the operative geographical units (OGUs), and the selection of a similarity coefficient or cluster method. Each of these decisions may lead to different results from the same data set.

A general methodology to carry out a quantitative biogeographical analysis uses a data set arranged as a matrix that scores the presence-absence of taxa (e.g., species) for each OGU selected (*see* Bricks 1987: fig. 1). The data matrix is transformed into a similarity (or dissimilarity) matrix to compare how similar (or different) each pair of OGUs are. Finally, the similarity matrix is used to carry out a cluster analysis, where one of several cluster algorithms must be chosen to generate a classification of the OGUs. The latter will help in proposing a geographical regionalization that supposedly has biological significance (Crovello 1981, McLaughlin 1986, Bricks 1987).

Different criteria to evaluate the similarity between pairs of OGUs, as seen in the similarity coefficients, will lead to different classifications. To what extent does the selection of one or another similarity index influence the resulting biogeographical classifications? This is a question that requires a response in view of the increasing number of analyses using the general methodology outlined above.

The literature documents the need to incorporate null models as a test of biogeographical hypotheses (Simberloff 1983, Craw 1989, Gotelli & Graves 1996). The null probability concept as the basis for constructing up null models has been discussed by Simberloff (1983), who has applied it to different biogeographical school methodologies (e.g. Simberloff 1978). Gotelli and Graves (1996) summarize the methods used to construct up null models.

The aim of this paper is to determine how the selection of a similarity coefficient affects resulting biogeographical classifications when using methods of quantitative biogeography.

## Methods

The classification of OGUs into higher level biogeographical units can be considered a process that selects one of many possible combinations. Naturally, the more OGUs being analyzed, the larger the number of possible combinations. In this paper we analyze with a combinatory approximation the different possible classifications resulting from the use of different similarity coefficients.

Two sources of information were used to investigate how the similarity coefficients affect biogeographic classifications. The first is a null model approach (Gotelli & Graves 1996) using an exhaustive set of hypothetical matrices of three OGUs and three, four, and five attributes (species or taxa), and comparing the classification topologies obtained after the cluster analysis. The second uses an empirical data set, a presence/absence data matrix of the genera of Mexican Asteraceae (368 genera), and the 32 political states (OGUs) of Mexico.

### Possible classifications

Part of the analysis was carried out on a set of all possible classifications of three OGUs. A presence/absence data matrix of three OGUs and three taxa can be ordered in 512 ($2^9$) possible combinations (each of the nine cells can be equally scored as an absence or a presence). Likewise, a data matrix of three OGUs and four taxa will have 4096 ($2^{12}$) possible combinations, and one with three OGUs and five taxa 32 768 ($2^{15}$) possible combinations.

Next, any of these equally possible datasets is transformed into a similarity matrix by using a similarity coefficient. This coefficient will transform the raw presence/absence data into values that measure how similar two OGUs are based on the number of taxa shared. The transformed similarity matrix is now a triangular matrix, because both halves contain the same data (Sneath & Sokal 1973).

Finally, the similarity matrix is used to construct a tree graph or dendrogram that will depict the relationships among the OGUs, based on the similarity values. This dendrogram (the topology used for the classification) is used to group the OGUs in higher level units that finally will be useful to propose a regionalization of the study area.

To compare the classifications obtained, we used the triplets method, similar to the quartet method (Estabrook *et al*. 1985). This method determines a criterion when a triplet has the same arrangement in two dendrograms. For a binary tree with $n$ terminal leaves (OGUs) there are $n(n-1)(n-2)/6$ triplets (Page 1993). Thus, for example the data matrix for the genera of Asteraceae of Mexico, which includes 32 OGUs, each classification includes 4960 triplets. The triplets method is more relaxed than that of quartets, and we consider it more appropriate in that it reduces overestimates of the differences among the classificatory topologies obtained in the dendrograms.

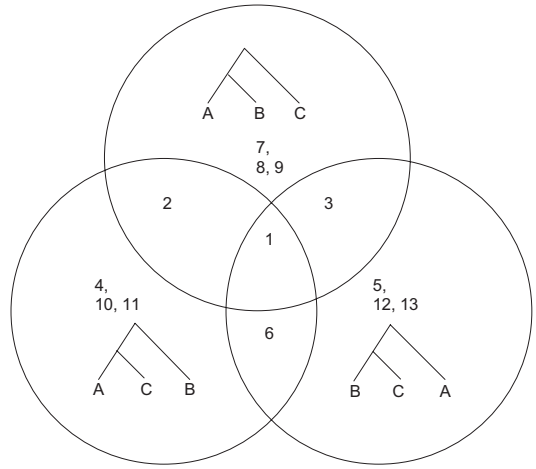


**Fig. 1.** Correspondence between matrix structures and classification topologies. The numbers correspond to matrix structures in Table 1.

## Matrix structures

In addition to analyzing the classification topologies (dendrograms) obtained from different similarity coefficients, the structure of the similarity matrices is likewise analyzed for the possible arrangement of values the matrices can produce (this arrangement is here referred to as the 'matrix structure', and can show equal, larger, or lower relationships among three pairs of OGUs). For example, a three-OGUs data matrix results in a similarity matrix that can produce 13 different equally possible hierarchical values (Table 1).

Each arrangement (matrix structure, Table 1) was then associated with one or more classification topologies, using a cluster algorithm (Fig. 1). Thus the most helpful structures to select the similarity coefficients that produce a single dendrogram were determined.

We employed the following rules to evaluate the arrangements of the similarity values in a data matrix:

a. Poorly informative structures: those with the three pairs of OGUs equal in values (AB = BC = AC) or those in which the similarity coefficient cannot be applied because the denominator is zero.
b. Half informative structures (enclosing structures): those that produce several dendrograms due to equal numerical values between two pair of OGUs in the similarity matrix (for example AB = AC > BC.)
c. Highly informative structures: structures with different values among the three pairs of OGUs that thus produce a single dendrogram (for example AB > AC > BC.)

The present study emphasizes the influence of the similarity coefficients in the construction of dendrograms. However, based on the former

**Table 1.** The 13 equally possible hierarchical values (matrix structures) that can produce a similarity matrix obtained from an original data matrix of three OGUs (A, B, and C) (=, > or < indicate the relative order of the similarity values, that is, equal, larger than, or lower than.) The structures are arranged according to their information content.

| |
|---|
| **Poorly informative structures:** |
| 1. AB = BC = AC |
| **Half informative structures:** |
| 2. AB = AC > BC |
| 3. AB = BC > AC |
| 4. AB = BC < AC |
| 5. AB = AC < BC |
| 6. BC = AC > AB |
| 7. BC = AC < AB |
| **Highly informative structures:** |
| 8. AB > AC > BC |
| 9. AB > BC > AC |
| 10. AC > BC > AB |
| 11. AC > AB > BC |
| 12. BC > AB > AC |
| 13. BC > AC > AB |

rules, it is possible to explore criteria for evaluating coefficients behavior.

It is important to point out that several matrix structures are associated with just one of the classification topologies (highly informative structures), while others can be associated with more than one (poorly or half informative structures) (Fig. 1). This arrangement is independent of the clustering method used (either single or complete linkages, or UPGMA).

For each data matrix, nine similarity matrices were produced by applying nine similarity coefficients (Table 2). Each similarity matrix was defined as poorly informative, half informative, or highly informative.

### Empirical analysis

In addition to evaluating the behavior of the similarity coefficients relative to the biogeographic classifications depicted as dendrograms based on theoretical matrices of three OGUs and three, four and five taxa, we conducted a similar analysis with empirical data: the genera of Asteraceae of Mexico. A presence/absence data matrix of 368 genera in the 32 political states of Mexico was transformed into nine similarity matrices using nine different coefficients (Table 2). Likewise, the dendrograms obtained from the use of three cluster algorithms (single linkage, complete linkage, and UPGMA) were analyzed. The analyses were made following the same procedure as with the theoretical data sets, that is, the dendrograms were compared with the triplets method to evaluate the differences resulting from the use of different similarity coefficients and different cluster algorithms, as well as the same coefficient but different cluster algorithms.

**Table 2.** Similarity coefficients used in the analysis.

| | |
|---|---|
| Simpson | $a/\min[(a+b),(a+c)]$ |
| Jaccard | $a/(a+b+c)$ |
| Braun-Blanquet | $a/\max[(a+b),(a+c)]$ |
| Sorensen-Dice | $a/[a+0.5(b+c)]$ |
| Kulczynski 1 | $a/(b+c)$ |
| Kulczynski 2 | $0.5[a/(a+b)+a/(a+c)]$ |
| Fager | $a/[(a+b)(a+c)]^{0.5}$ $-0.5\{\max[(a+b),(a+c)]\}$ |
| Otsuka | $a/[(a+b)(a+c)]^{0.5}$ |
| Correlation ratio | $a^2/[(a+b)(a+c)]$ |

$a$ = number of taxa (attributes) present in both OGUs
$b$ = number of taxa present only in the first OGU
$c$ = number of taxa present only in the second OGU

## Results

Table 3 shows the percentages of matrix structures generated with each data matrix analyzed using the nine similarity coefficients cited in Table 2. In the three data matrices, the coefficients with the largest percentages of poorly

**Table 3.** Percentages of structures, arranged according to their information content, obtained from the use of nine similarity coefficients in three different group matrices (Poorly = poorly informative; Half = Half informative; Highly = Highly informative).

| Similarity coefficient | Poorly | Half | Highly |
|---|---|---|---|
| **Three taxa three OGUs group matrices** | | | |
| Simpson | 50.6 | 42.4 | 7.0 |
| Jaccard | 15.0 | 67.4 | 17.6 |
| Braun-Blanquet | 18.5 | 70.9 | 10.5 |
| Sorensen-Dice | 15.0 | 67.4 | 17.6 |
| Kulczynski 1 | 27.3 | 55.1 | 17.6 |
| Kulczynski 2 | 57.8 | 35.2 | 7.0 |
| Fager | 36.7 | 45.7 | 17.6 |
| Otsuka | 36.7 | 45.7 | 17.6 |
| Correlation ratio | 43.7 | 49.2 | 7.0 |
| **Four taxa three OGUs group matrices** | | | |
| Simpson | 34.8 | 22.1 | 43.1 |
| Jaccard | 8.4 | 14.5 | 77.1 |
| Braun-Blanquet | 12.8 | 20.6 | 66.6 |
| Sorensen-Dice | 8.4 | 14.5 | 77.1 |
| Kulczynski 1 | 13.2 | 17.7 | 69.1 |
| Kulczynski 2 | 37.3 | 22.1 | 40.6 |
| Fager | 21.2 | 24.7 | 54.1 |
| Otsuka | 21.2 | 14.5 | 64.3 |
| Correlation ratio | 26.6 | 18.6 | 54.8 |
| **Five taxa three OGUs group matrices** | | | |
| Simpson | 23.0 | 24.7 | 52.3 |
| Jaccard | 7.5 | 13.1 | 82.4 |
| Braun-Blanquet | 8.3 | 20.9 | 70.9 |
| Sorensen-Dice | 7.5 | 13.1 | 82.4 |
| Kulczynski 1 | 6.2 | 14.3 | 79.5 |
| Kulczynski 2 | 22.1 | 23.6 | 54.3 |
| Fager | 11.6 | 21.9 | 66.4 |
| Otsuka | 11.6 | 14.7 | 73.6 |
| Correlation ratio | 14.4 | 20.5 | 65.1 |

informative structures generated were Simpson and Kulczynski 2. On the other hand, the coefficients that generated in the three data matrices the largest percentages of highly informative structures were Jaccard and Sorensen-Dice.

Because different classification topologies may be produced by one similar structure, it is important to point out that the choice of a coefficient will affect qualitatively the results of the biogeographical analysis. Only two coefficients (Jaccard and Sorensen-Dice) produced identical results (matches) both in the matrix structures and the classification topologies obtained. Those with lower percentages of matches were Fager and Braun-Blanquet with 38.0%, 36.1% and 39.2% of identical matrix structures for three, four and five taxa matrices, and 50.2%, 56.0% and 60.5% of identical triplets for three, four and five taxa matrices. The results point out that two equally feasible biogeographical classifications may be obtained from the same data matrix by using different similarity coefficients, ranging from 50.2% to 100%.

The analysis of the genera of Asteraceae of Mexico also showed contrasting results, parallel to those obtained from the theoretical analysis. Table 4 compares the classification topologies obtained from the use of three different clustering methods. Percentages of similar topologies showed differences among the coefficients used. They go from 29.5% (Braun-Blanquet-Correlation Ratio, single linkage) to 100% (for example

**Table 4.** Percentages of identical classification topologies (triplets) obtained from the use of nine different similarity coefficients and three cluster methods, applied to a presence/absence data matrix of the genera of Asteraceae (368) occurring in the states of Mexico (32).

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| **Complete linkage** | | | | | | | | | |
| (1) Simpson | 100.0 | | | | | | | | |
| (2) Jaccard | 67.5 | 100.0 | | | | | | | |
| (3) Braun-Blanquet | 60.6 | 87.4 | 100.0 | | | | | | |
| (4) Sorensen-Dice | 67.5 | 100.0 | 87.4 | 100.0 | | | | | |
| (5) Kulczynski 1 | 67.5 | 100.0 | 87.4 | 100.0 | 100.0 | | | | |
| (6) Kulczynski 2 | 82.4 | 73.7 | 69.8 | 73.7 | 73.7 | 100.0 | | | |
| (7) Fager | 68.5 | 61.2 | 56.0 | 61.2 | 61.2 | 77.6 | 100.0 | | |
| (8) Otsuka | 67.5 | 100.0 | 87.8 | 100.0 | 100.0 | 73.7 | 61.2 | 100.0 | |
| (9) Correlation ratio | 71.2 | 66.6 | 62.7 | 66.6 | 66.6 | 71.6 | 66.4 | 66.6 | 100.0 |
| **Single linkage** | | | | | | | | | |
| (1) Simpson | 100.0 | | | | | | | | |
| (2) Jaccard | 42.6 | 100.0 | | | | | | | |
| (3) Braun-Blanquet | 35.4 | 79.7 | 100.0 | | | | | | |
| (4) Sorensen-Dice | 42.6 | 100.0 | 79.7 | 100.0 | | | | | |
| (5) Kulczynski 1 | 68.9 | 100.0 | 79.7 | 100.0 | 100.0 | | | | |
| (6) Kulczynski 2 | 67.8 | 52.0 | 46.6 | 52.0 | 52.0 | 100.0 | | | |
| (7) Fager | 41.7 | 39.1 | 30.1 | 39.1 | 39.1 | 53.6 | 100.0 | | |
| (8) Otsuka | 54.2 | 94.9 | 81.9 | 94.9 | 94.9 | 55.0 | 37.1 | 100.0 | |
| (9) Correlation ratio | 100.0 | 29.8 | 29.5 | 29.8 | 29.8 | 50.4 | 31.5 | 32.4 | 100.0 |
| **UPGMA** | | | | | | | | | |
| (1) Simpson | 100.0 | | | | | | | | |
| (2) Jaccard | 61.0 | 100.0 | | | | | | | |
| (3) Braun-Blanquet | 55.7 | 72.1 | 100.0 | | | | | | |
| (4) Sorensen-Dice | 61.1 | 100.0 | 72.1 | 100.0 | | | | | |
| (5) Kulczynski 1 | 54.5 | 83.6 | 66.2 | 83.6 | 100.0 | | | | |
| (6) Kulczynski 2 | 74.8 | 72.3 | 64.0 | 77.3 | 69.6 | 100.0 | | | |
| (7) Fager | 74.7 | 66.2 | 50.2 | 66.2 | 55.9 | 75.4 | 100.0 | | |
| (8) Otsuka | 62.7 | 88.1 | 72.7 | 88.1 | 78.3 | 69.5 | 56.7 | 100.0 | |
| (9) Correlation ratio | 62.5 | 45.6 | 52.2 | 45.6 | 39.4 | 53.4 | 59.7 | 49.5 | 100.0 |

Jaccard–Sorensen-Dice in the three clustering methods).

Table 5 shows the percentages of similar classification topologies of Table 4, comparing the same similarity coefficient but differing in clustering method. In all cases similar topologies never matched totally; values ranged from 51.0 (Complete linkage *vs*. Single linkage, based on Fager) to 84.1% (Complete linkage *vs*. UPGMA, based on Jaccard and Sorensen-Dice). The Complete Linkage and UPGMA methods produced the greatest number of matching classification topologies, no matter the similarity coefficient used (Table 5). The Complete and Single Linkage methods produced the lesser matching classification topologies. The differences are expected because the UPGMA method is intermediate among the Complete Linkage and the Single Linkage methods, which use respectively the maximum and minimum similarity values.

## Discussion and conclusion

The greatest number of similar classification topologies were obtained from the use of the Jaccard and Sorensen-Dice similarity coefficients. They also showed the least number of poorly informative structures. The best behavior of these coefficients agree with the results presented by Hubálek (1982), who concluded that these are the best coefficients based on a series of specific criteria.

**Table 5.** Percentages of identical classification topologies (triplets) obtained with the three cluster methods applied to a presence/absence data matrix of the genera of Asteraceae (368) occurring in the states of Mexico (32).

| | Complete *vs.* Single | Complete *vs.* UPGMA | Single *vs.* UPGMA |
|---|---|---|---|
| Simpson | 52.7 | 68.2 | 60.5 |
| Jaccard | 66.2 | 84.1 | 71.9 |
| Braun-Blanquet | 69.7 | 79.9 | 73.3 |
| Sorensen-Dice | 66.2 | 84.1 | 71.9 |
| Kulczynski 1 | 66.2 | 78.3 | 76.6 |
| Kulczynski 2 | 55.2 | 73.4 | 61.1 |
| Fager | 51.0 | 73.0 | 58.5 |
| Otsuka | 64.8 | 77.7 | 69.2 |
| Correlation ratio | 47.1 | 64.7 | 62.3 |

Our results do not agree with that of Sánchez and López (1988), who concluded that Simpson's coefficient was adequate for biogeographical studies. This coefficient, along with Kulczynski 2 produced lower percentages of highly informative structures among nine similarity coefficients analyzed (Table 3). Also, they showed the lowest percentage values when applied to the empirical data (Tables 4 and 5). Based on our results, it is surprising that Hubálek (1982) included Kulczynski 2 as a "good" coefficient, along with Jaccard and Sorensen-Dice.

Our analysis of empirical data resulted in classificatory topologies as different as 51% (Table 4), depending on the clustering method used. The results indicate that care should be used in selecting a clustering method. The clustering method definitely influences strongly the classification's topology.

Although there are no solid arguments in favor of a particular similarity coefficient, the analyses of their behavior and properties (*see* also Hubálek 1982, Sánchez & López 1988) help guide the choice of a coefficient. However, if several coefficients are to be used, additional criteria such as those evaluated in this paper can be used. Our results suggest that based on the number of poorly informative structures generated, the best index to use is Jaccard or Sorensen-Dice, followed by Braun-Blanquet or Kulczynski 1.

The biogeographical classifications produced by clustering methods rely strongly on the similarity coefficient and the clustering methods used. Thus, as our results indicate, the classifications obtained from different data sets employing different similarity coefficients and clustering methods are not comparable. These classifications obtained by quantitative methods, should be used as heuristic guides to define biogeographical regions. They must be compared with alternative forms of analysis, for example, using phylogenetic principles as in PAE analyses (Rosen 1988) or panbiogeographic proposals (Craw 1989) to avoid relying only on a single interpretation.

In this paper we use null models design principles in similarity coefficients and clustering methods, in an equivalent way to the equiprobable cladograms of Simberloff (1983). As did

Simberloff, we quantified the differences and probabilities shown by similar methods, independent of the data used in a biogeographical study. Our analysis could be considered an additional null model that explores the properties, goodness and failures of different similarity coefficients and clustering methods to biogeographical presence/absence data.

## Acknowledgements

The senior author acknowledges the Dirección General de Asuntos del Personal Académico of the Universidad Nacional Autónoma de México, the economic support through a doctoral fellowship. The careful reading by Drs. Claudio Delgadillo, Luis Eguiarte, Jorge Llorente, and Mark Olson improved the ideas and content of the paper.

## References

Bricks, H. J. B. 1987: Recent methodological developments in descriptive biogeography. — *Ann. Zool. Fennici* 24: 165–178.

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

Craw, R. 1989: Quantitative panbiogeography: introduction to methods. — *New Zealand J. Zool.* 16: 485–494.

Crovello, T. J. 1981: Quantitative biogeography: an overview. — *Taxon* 30: 563–575.

Estabrook, G. F., McMorris, F. R. & Meacham, C. A. 1985: Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. — *Syst. Zool.* 34: 193–200.

Gotelli, N. J. & Graves, G. R. 1996: *Null models in ecology*. — Smithsonian Inst. Press, Washington, D.C.

Hubálek, Z. 1982: Coefficients of association and similarity, based on binary (presence-absence) data: an evaluation. — *Biol. Rev.* 57: 669–689.

McLaughlin, S. P. 1986: Floristic analysis of the southwestern Unites States. — *Great Basin Naturalist* 46: 46–65.

Page, R. D. M. 1993: *Component. Version 2.0. User's guide*. — Nat. Hist. Mus., London.

Rosen, B. R. 1988: From fossils to earth history: applied historical biogeography. — In: Myers, A. A. & Giller, P. S. (eds.), *Analytical biogeography*: 437–481. Chapman & Hall, New York.

Sánchez, O. & López, G. 1988: A theoretical analysis of some indices of similarity as applied to biogeography. — *Folia Entomológica Mexicana* 75: 119–145.

Simberloff, D. 1978: Using island biogeographic distributions to determine if colonization is stochastic. — *Am. Nat.* 112: 713–726.

Simberloff, D. 1983: Biogeography: the unification and maturation of a science. — In: Brush, A. H. & Klark, G. H. (eds.), *Perspectives in ornithology*: 411–455. Cambridge Univ. Press, Cambridge.

Sneath, P. H. A. & Sokal, R. R. 1973: *Numerical taxonomy*. — W. H. Freeman, San Francisco.