Rev. Brasil. Genet. 16, 3, 829-839 (1993)
(Brazil. J. Genetics)

METHODOLOGY

HIERARQUICAL VERSUS NONHIERARQUICAL PATTERNS OF GENETIC DISTANCES AMONG POPULATIONS: A SIMULATION STUDY

José Alexandre Felizola Diniz-Filho

ABSTRACT

A simulation study was made of the effects of mixing two evolutionary forces (natural selection and random genetic drift), combined in a single data matrix of gene frequencies, on the resulting genetic distances among populations. Twenty-one kinds of simulated gene frequencies surfaces, for 15 populations linearly distributed over geographic space, were used to construct 21 data matrices, combining different proportions of two types of surfaces (gradients and random surfaces). These matrices were analysed by Unweighted Pair-Group Method - Arithmetic Averages (UPGMA), clustering and Principal Coordinate Analysis. The results obtained show that ordination is more accurate than UPGMA in revealing the spatial patterns in the genetic distances, in comparison with results obtained using the Mantel test comparing directly genetic and geographic distances.

INTRODUCTION

Cluster analysis is a method for representing a similarity matrix among taxonomic unities by a treelike arrangement called a dendrogram (Romesburg, 1984). Most clustering methods used in biological research operate with sequential, aglomerative, hierarquical and non-overlapping (SAHN) algorithms (Sneath and Sokal, 1973). The agreement between the classification obtained with the dendrogram and the original distance matrix can be evaluated (though not in an statistical sense) by the cophenetic correlation coefficient (Sokal and Rohlf, 1962).

Departamento de Biologia, Instituto de Biociências, UNESP, Caixa Postal 199, 13506-900 Rio Claro, SP, Brasil.

Clustering techniques have been widely used in population genetics to produce an hierarquical arrangement of populations, based upon Nei's (1972, 1978) or Rogers's (1972) genetic distances among them (Buth, 1984). The most commonly used SAHN technique is the UPGMA (unweighted pair-group method - arithmetic average) (Sneath and Sokal, 1973). Classifications produced by this technique possess a number of desirable properties, such as stability and maximization of the cophenetic correlation coefficient (Farris, 1969; Rohlf and Sokal, 1981). UPGMA based upon genetic distances has also been considered an efficient estimator of cladogeny (Nei *et al.*, 1983; Sokal, 1986a).

Although UPGMA clustering is the best of the SAHN techniques, its application subsumes the existence of an hierarquical structure of similarity among taxonomic unities (populations), and this situation is not always found at the populational level (Sokal *et al.*, 1987). Different evolutionary factors, such as natural selection, migration and genetic drift, can affect gene frequencies distributed over geographic space (gene frequencies surfaces) (Sokal and Oden, 1978a,b), producing different pattens of spatial variation (Sokal, 1978; Manly, 1985). The combination of these different patterns should not always produce an hierarquical multivariate structure of genetic similarity among populations. This problem must be reflected in the dendrograms obtained with UPGMA clustering, producing low values of the cophenetic correlation coefficient, indicating distortions in representing the multivariate space.

When different gene frequencies are highly correlated over geographic space, they must be under the effect of the same selective agent, producing some kind of spatial pattern, such as patches or clines (if there is no linkage among them) (Sokal and Wartenberg, 1981; Sokal *et al.*, 1989). Considering that gene frequency surfaces produced by genetic drift are not expected to be correlated over geographic space (Sokal and Wartenberg, 1983; Sokal, 1986b), when we increase the number of random gene frequency surfaces in a data matrix, we are reducing the power of natural selection to produce a clear multivariate spatial pattern of genetic similarity among populations.

In this paper, we analysed simulated data sets, trying to understand the effects of combining these two kinds of gene frequencies surfaces (gradients and random surfaces) on the multivariate statistical analyses of genetic distances among populations.

MATERIAL AND METHODS

We simulated a geographic map containing 15 populations, linearly distributed over space. For these populations, two kinds of gene frequencies surfaces were generated: (1) a cline, assumed here to be produced by natural selection, with the following values of gene frequencies: 0.28, 0.31, 0.32, 0.35, 0.43, 0.43, 0.44, 0.53, 0.55, 0.56, 0.57, 0.63, 0.65, 0.67, 0.72 (surface of type A); (2) random arrangements of these values over the same geographic space (surfaces of type B). Twenty different surfaces of type B were

produced, by a random arrangement of each value of gene frequency to each population. These random surfaces are here assumed to be due to random genetic drift within populations, producing no spatial pattern among them (Sokal and Oden, 1978b). The presence or absence of spatial pattern on these 20 surfaces was checked with Moran's I spatial autocorrelation coefficient (Sokal and Oden, 1978a).

With the 21 surfaces, we constructed 21 data matrices, combining different proportions of the two kinds of gene frequency surfaces, as follows: 20A+OB, 19A+1B, 18A+2B, and so on. Each data matrix, therefore, had 20 gene frequency surfaces and 15 populations. With this procedure, we created different contrasts of surfaces, in a single data matrix, produced by two evolutionary forces: (1) natural selection, producing a cline of highly correlated surfaces; (2) random genetic drift, producing random variation of the gene frequencies over the geographic space (Barbujani, 1987).

The 21 data matrices were used to obtain Rogers (1972) genetic distances among populations. The choice of this coefficient was based on its metric properties, desirable when using clustering and ordination techniques (Nei *et al.*, 1983). Rogers genetic distance estimates the mean geometric distance between gene frequencies, summarizing this information across all loci (Buth, 1984).

The genetic distances among populations were clustered using UPGMA (Sneath and Sokal, 1973), producing dendrograms which were tested by the cophenetic correlation coefficient (R_c). The same distances were also submitted to a Principal Coordinate Analysis (PCORD) (Gower, 1966). The coordinates of each population in the first three eigenvectors were used to construct a matrix of Euclidean distances among them, which was compared with the original genetic distance matrix. The matrix correlation thus obtained (R_o) permits to check the magnitude of distortion in the genetic distances by the ordination in a reduced multivariate space (Rohlf, 1978; Rohlf and Sokal, 1981).

The relationships between multivariate arrangements of populations (by clustering and ordination) and the patterns of geographical variability were analysed using the Mantel test (Sokal, 1979). The 21 original genetic distances, the 21 cophenetic matrices and the 21 Euclidean distances among populations in the reduced multivariate space were compared with the matrix of geographic distances among populations using the Mantel test. The statistical significance of the normalized Mantel's Z (= matrix correlation) (Smouse *et al.*, 1986) was obtained with 1000 random permutations of one of the matrices (Manly, 1991). So, it was possible to analyse the capacity of the multivariate methods in detecting spatial patterns of genetic variability among populations, using the matrix correlations between original genetic distances and geographic distances as a non-disturbed reference.

All the multivariate statistical analyses and Mantel tests were performed with a PC/AT microcomputer, using the Numerical Taxonomy and Multivariate Analysis System (NTSYS-PC), version 1.5 (Rohlf, 1989).

RESULTS

Examples of the two kinds of surfaces generated can be seen in Figure 1. The results of spatial autocorrelation analysis for each of the 21 surfaces simulated are in Table I. It is possible to see from the standard normal deviation (SND) of each Moran's I coefficient that only the surface of type A has a significant ($P < 0.01$) spatial autocorrelation, confirming the absence of spatial pattern in the surfaces of type B. This is a basic condition for the subsequent analyses.

Table I - Moran's I (I) and its standard normal deviate (SND) obtained in the spatial autocorrelation analysis of the 21 simulated surfaces.

Surface	Type	I	SND
1	A	0.444**	5.916
2	B	0.059^{ns}	1.643
3	B	-0.120^{ns}	0.621
4	B	0.088^{ns}	1.955
5	B	0.006^{ns}	0.990
6	B	-0.152^{ns}	0.974
7	B	-0.049^{ns}	0.267
8	B	-0.117^{ns}	0.590
9	B	-0.156^{ns}	1.048
10	B	-0.162^{ns}	1.146
11	B	-0.152^{ns}	1.019
12	B	-0.093^{ns}	0.272
13	B	-0.084^{ns}	0.155
14	B	-0.141^{ns}	0.841
15	B	-0.047^{ns}	0.307
16	B	0.070^{ns}	0.013
17	B	0.159^{ns}	1.089
18	B	-0.100^{ns}	0.354
19	B	-0.113^{ns}	0.533
20	B	-0.200^{ns}	1.562
21	B	0.004^{ns}	0.995

**, $P < 0.01$; ns, nonsignificant (5%).

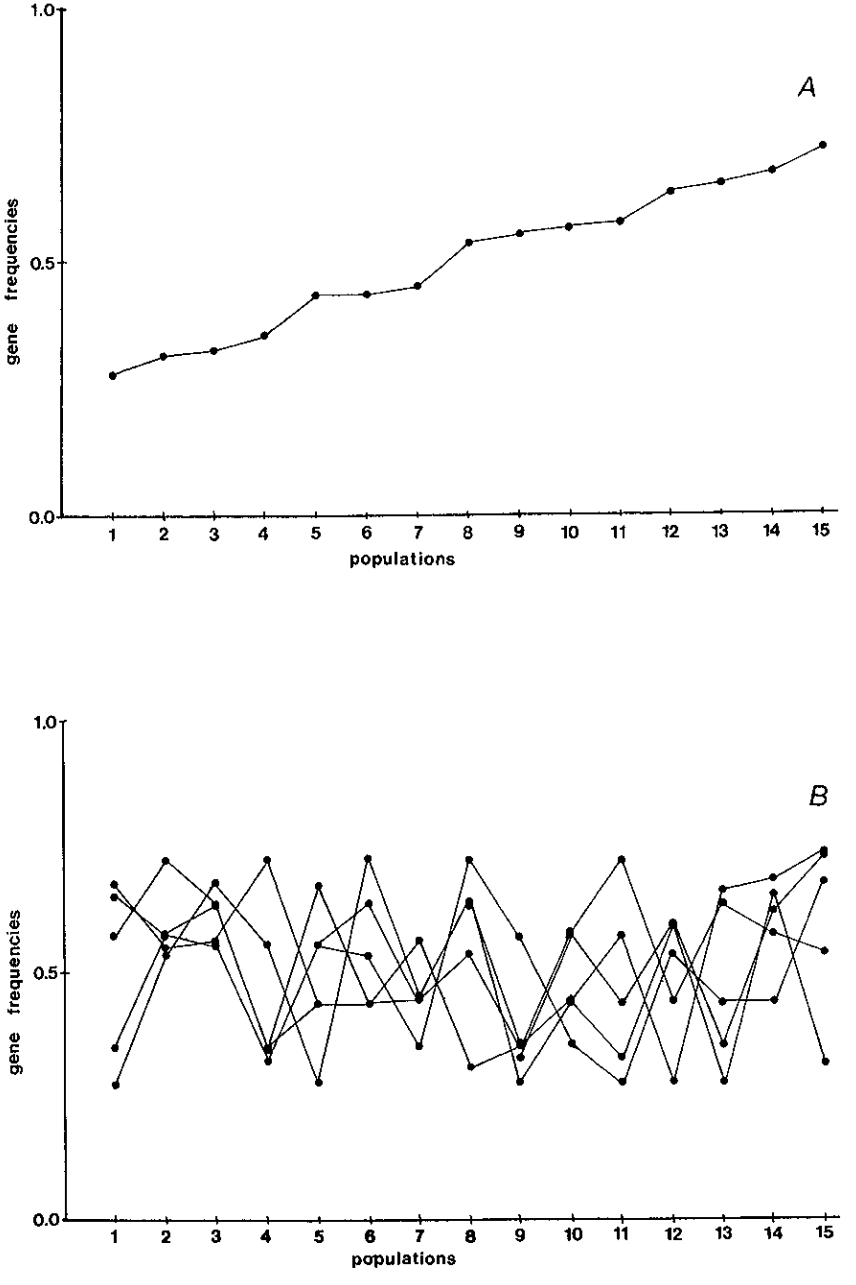

Figure 1 - Examples of the two types of gene frequencies surfaces analysed. A) Surface of type A (gradient); B) Surfaces of type B (random surfaces). Populations ordered by geographic location.

Table II shows the results of the analysis in the 21 genetic distances matrices. The cophenetic correlation coefficients of the UPGMA clustering have low values, ranging from 0.6684 to 0.5608, considerably below the value established by Sneath and Sokal (1973) and Sokal (1986a) for a "good" representing of the multivariate similarity structure (0.85). The decrease in the cophenetic values is also highly correlated with the increase in the number of random gene frequencies surfaces in the data matrices (Spearman's $r_S = -0.872$; $P < 0.01$).

Table II - Obtained data matrices, number of random gene frequency surfaces (NRS) in data matrix and results of UPGMA clustering and PCORD.

Data matrix	NRS	R_c	λ_1	R_o
A20+B00	0	0.631	15.360	0.844
A19+B01	1	0.668	14.335	0.844
A18+B02	2	0.667	13.212	0.848
A17+B03	3	0.668	12.181	0.855
A16+B04	4	0.663	11.298	0.853
A15+B05	5	0.659	10.347	0.860
A14+B06	6	0.662	9.344	0.866
A13+B07	7	0.659	8.435	0.869
A12+B08	8	0.649	7.652	0.876
A11+B09	9	0.641	6.811	0.876
A10+B10	10	0.615	5.952	0.893
A09+B11	11	0.609	5.222	0.870
A08+B12	12	0.601	4.580	0.852
A07+B13	13	0.586	3.997	0.844
A06+B14	14	0.561	3.527	0.919
A05+B15	15	0.566	2.949	0.895
A04+B16	16	0.566	2.508	0.881
A03+B17	17	0.592	2.192	0.857
A02+B18	18	0.566	1.988	0.809
A01+B19	19	0.569	1.892	0.824
A00+B20	20	0.568	1.809	0.797

The results of PCORD are also given in Table II. The values of the first eigenvalue obtained (λ_1) are clearly associated with changes in the number of random

gene frequencies surfaces in the data matrices, which was expected when considering the reduction in the correlations among the variables (gene frequencies surfaces) in the data matrices. The matrix correlations between original genetic distances and distances in the multivariate reduced space (R_o) are considerable higher than the cophenetic correlations, ranging from 0.919 to 0.797. These values are not correlated with changes in the number of gene frequencies surfaces in the data matrices (r_s = -0.029; P > 0.05).

The normalized Z values obtained with the Mantel test are in Table III. The decrease in the coefficients is apparently related to changes in the number of gene frequencies surfaces in the data matrices. However, almost all of them are statistically significant, indicating the presence of some spatial pattern in the genetic distances.

Table III - Normalized Z values obtained with the Mantel test.

Data matrix	Z1	Z2	Z3
A20+B00	0.948**	0.648**	0.867**
A19+B01	0.945**	0.706**	0.871**
A18+B02	0.944**	0.705**	0.875**
A17+B03	0.942**	0.704**	0.871**
A16+B04	0.938**	0.706**	0.862**
A15+B05	0.934**	0.703**	0.865**
A14+B06	0.929**	0.695**	0.869**
A13+B07	0.921**	0.696**	0.881**
A12+B08	0.913**	0.688**	0.880**
A11+B09	0.905**	0.688**	0.880**
A10+B10	0.893**	0.573**	0.881**
A09+B11	0.874**	0.528**	0.872**
A08+B12	0.836**	0.658**	0.822**
A07+B13	0.784**	0.492**	0.768**
A06+B14	0.723**	0.393**	0.723**
A05+B15	0.652**	0.344**	0.666**
A04+B16	0.562**	0.226**	0.596**
A03+B17	0.456**	0.219*	0.488**
A02+B18	0.287*	0.086^{ns}	0.257*
A01+B19	0.147^{ns}	0.059^{ns}	0.140^{ns}
A00+B20	-0.050^{ns}	-0.119^{ns}	0.025^{ns}

**, P < 0.01; *, 0.01 < P < 0.05; ns, nonsignificant (5%).

Figure 2 shows the relationships between Z values and number of random gene frequency surfaces in the data matrices. The Z values for the comparison between original genetic distances and geographic distances (Z1) can be used as a reference curve, since there is no distortion produced by multivariate representation of relationships among populations. The Z values obtained by comparing cophenetic matrices and geographic distances (Z2) were always lower than the reference Z values, which can be explained by distortion in the patterns of genetic distances produced by clustering. More importantly, the Z values decrease faster then the reference ones, indicating that clustering reduces the visualization of some spatial pattern in the distance matrices. Among the Z2 values, the last three Z values were nonsignificant, indicating that even when there are two clines in the data matrix, no spatial pattern of genetic distances are revealed. Z values for comparison of geographic distances and Euclidean distances among populations in the multivariate reduced space (Z3) were very similar to reference Z values, indicating that ordination preserves spatial patterns in the genetic distances with very small distortion. Only the two last values of Z3 gave no significant Z values, indicating that even when there are just two clines in the data matrix, the ordination indicates some spatial pattern in the genetic distances, as in the reference Z values.

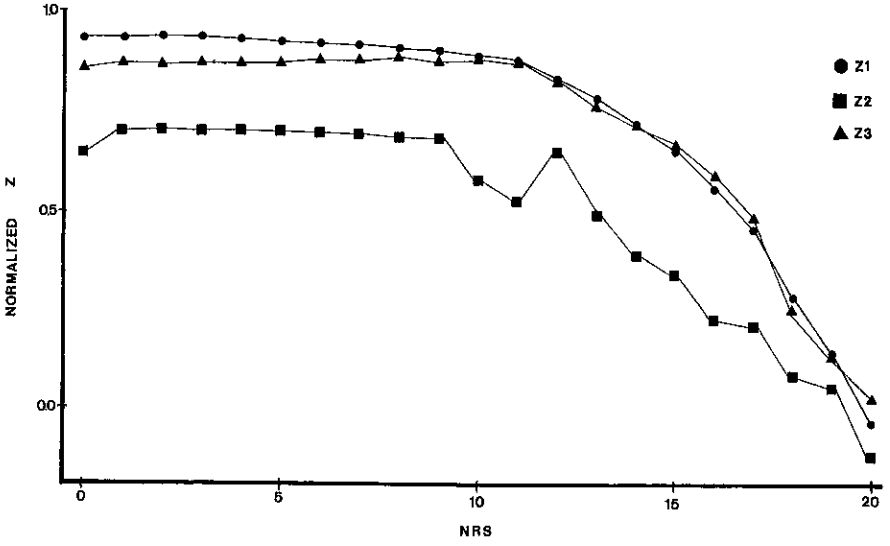

Figure 2 - Relationship between normalized Z values for the three matrices compared (original genetic distances, cophenetic matrices and Euclidean distances in the ordination reduced space) and number of random gene frequencies surfaces in the data matrix (NRS).

DISCUSSION

The results show that UPGMA clustering produces more distortions in the genetic distances than ordination techniques. More importantly, changes in the cophenetic correlation coefficients are associated with changes in the number of random gene frequencies surfaces in the data matrix. This situation must be frequently found in real data sets, in which different evolutionary forces determine different spatial patterns in the gene frequencies. The cophenetic values obtained with UPGMA were always very low, indicating that even when all gene frequencies are patterned as clines over geographic space, there is not a good representation, by clustering, of genetic distances. Of course, in this situation, hierarquical arrangements are not adequate. So, when there is a mixture of random gene frequency surfaces and gradients over space, clustering methods are not adequate to analyse genetic distances. It is much better to represent these distances in a continuous multivariate space, produced by ordination techniques, such as PCORD.

The values of R_o obtained were considerably higher than the cophenetic values, indicating that there is less distortion in this multivariate representation. More importantly, these values were not correlated with the number of random gene frequencies surfaces in the data matrix. We understand that the eigenanalysis of double centered genetic distances, which is the basis of PCORD (Rohlf, 1972), preserves clines in principal axes even when there is a great number of random surfaces in the data matrix, keeping the original spatial pattern of genetic distances, as revealed by the Mantel test applied to the results of PCORD (Figure 2). This can be an advantage in ordering multivariate patterns of genetic variation, since random surfaces are not adequate to do so, being considered as residuals in these overall patterns.

So, we conclude that the establishment of similarity patterns, using genetic distances among populations, should be carefully done when using UPGMA clustering, and that cophenetic correlation coefficients must always be computed. Ordination techniques give better results anyway. When there is a great proportion of random gene frequency surfaces in the data matrix, ordinations should be used instead, by considering their property to keep patterned surfaces in the principal axes. It is clear that the choice of a multivariate technique to represent genetic distances among populations must be conditioned by the types of spatial patterns in the gene frequencies.

This paper presents a very simple simulated situation, contrasting a single selective agent, patterned as a cline, with random gene frequency surfaces, produced by genetic drift. Real situations must be much more complex, with other kinds of surfaces produced by distinct selective agents and evolutionary forces. This increase in complexity must affect the results of multivariate representation of genetic distances in a much more complex way.

ACKNOWLEDGMENTS

Thanks are due to Paulo de Marco Junior, Maurício Bini, Maria Izabel B. Pignata, Regina Bessi and to an anonymous reviewer for helpful suggestions made in earlier drafts of this manuscript. The author is recipient of a CNPq Master's degree Fellowship.

Publication supported by FAPESP.

RESUMO

Um estudo simulado foi realizado a fim de compreender os efeitos de combinar dois tipos de superfícies de freqüências gênicas (produzidas por seleção natural e deriva genética) sobre as distâncias genéticas. Vinte e uma superfícies de freqüências gênicas foram simuladas, para 15 populações linearmente distribuídas sobre o espaço geográfico, sendo então utilizadas para gerar 21 matrizes de distância genética, combinando diferentes proporções dos dois tipos de superfície. Essas matrizes foram analisadas por Análises de Agrupamento (UPGMA) e Coordenadas Principais. Os resultados mostram que a ordenação é mais acurada no sentido de revelar os padrões espaciais das freqüências gênicas existentes nas distâncias genéticas, considerando os valores obtidos nos testes de Mantel aplicados na comparação de matrizes de distância genética (originais e reduzidas pelos métodos de análise) e geográfica entre as populações.

REFERENCES

Barbujani, G. (1987). Autocorrelation of gene frequencies under isolation-by-distance. *Genetics 177*: 772-782.

Buth, D.G. (1984). The application of electrophoretic data in systematic studies. *Ann. Rev. Ecol. Syst. 15*: 501-522.

Farris, J. (1969). On the cophenetic correlation coefficient. *Syst. Zool. 18*: 279-285.

Gower, J.C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika 53*: 325-338.

Manly, B.F.J. (1985). *The Statistics of Natural Selection*. Chapman & Hall, London.

Manly, B.F.J. (1991). *Randomization and Monte Carlo Methods in Biology*. Chapman & Hall, London.

Nei, M. (1972). Genetic distances among populations. *Am. Nat. 106*: 283-292.

Nei, M. (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics 89*: 583-590.

Nei, M., Tajima, F. and Tateno, Y. (1983). Accuracy of estimated phylogenetic trees from molecular data. II. Gene frequency data. *J. Mol. Evol. 19*: 153-170.

Rogers, J.S. (1972). Measures of genetic similarity and genetic distance. *Univ. Tex. Publ. 7213*: 145-153.

Rohlf, F.J. (1972). An empirical comparison of three ordination techniques in numerical taxonomy. *Syst. Zool. 21*: 271-280.

Rohlf, F.J. (1978). Methods of comparing classifications. *Ann. Rev. Ecol. Syst. 5*: 101-113.

Rohlf, F.J. (1989). *NTSYS-Pc: Numerical Taxonomy and Multivariate Analysis System*. Exeter, New York.

Rohlf, F.J. and Sokal, R.R. (1981). Comparing numerical taxonomic studies. *Syst. Zool. 30*: 459-490.

Romesburg, H.C. (1984). *Cluster Analysis for Researchers*. Wadsworth, London.

Smouse, P.E., Long, J.C. and Sokal, R.R. (1986). Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Syst. Zool. 35*: 627-632.

Sneath, P.H.A. and Sokal, R.R. (1973). *Numerical Taxonomy*. W.H. and Freeman, San Francisco.

Sokal, R.R. (1978). Population differentiation: Something new or more of the same? In: *Genetics and Ecology* (Brussard, P. and Solbrig, O., eds.). Academic Press, New York.

Sokal, R.R. (1979). Testing statistical significance of geographic variation patterns. *Syst. Zool. 28*: 227-232.

Sokal, R.R. (1986a). Phenetic taxonomy: theory and methods. *Ann. Rev. Ecol. Syst. 17*: 423-442.

Sokal, R.R. (1986b). Spatial data analysis and historical processes. In: *Data Analysis and Informatics IV* (Diday *et al.*, eds.). Elsevier Science Publishers, Holland.

Sokal, R.R. and Oden, N.L. (1978a). Spatial autocorrelation in biology. 1. Methodology. *Biol. J. Linn. Soc. 10*: 199-228.

Sokal, R.R. and Oden, N.L. (1978b). Spatial autocorrelation in biology. 2. Some biological implications and four applications of evolutionary and ecological interest. *Biol. J. Linn. Soc. 10*: 229-249.

Sokal, R.R. and Rohlf, F.J. (1962). The comparison of dendrograms by objective methods. *Taxon 9*: 33-40.

Sokal, R.R. and Wartenberg, D. (1981). Space and population structure. In: *Dynamic Spatial Models* (Griffith, D. and McKinnon, R., eds.). Sijthoff and Noordhoff, Netherlands.

Sokal, R.R. and Wartenberg, D. (1983). A test of spatial autocorrelation using an isolation-by-distance model. *Genetics 105*: 219-237.

Sokal, R.R., Uytterschaut, H., Rosing, F. and Schwidetzky, I. (1987). A classification of European skulls from three time periods. *Am. J. Phys. Anthr. 74*: 1-20.

Sokal, R.R., Jacquez, G.M. and Wooten, M.C. (1989). Spatial autocorrelation analysis of migration and selection. *Genetics 121*: 845-855.

(Received October 31, 1991)