KOREAN JOURNAL OF ZOOLOGY
Vol. 24, No. 2, 91-98 (1981)

Comparison of Reseults using Average Taxonomic Distance and Correlation Coefficient Matrices for Cluster Analyses

Hung Sun Koh
(Department of Biology, Chungbuk University)

Cluster Analyses에서 Average Taxonomic Distance와 Correlation Coefficient 행렬식들을 이용한 결과의 비교

고 홍 신
(충북대학교 생물학과)

(Received March 2, 1981)

적 요

Deer mice, *Peromyscus maniculatus*, 의 성체 571마리의 30개 morphometric 형질들을 이용한 cluster analyses에서 두가지의 similarity행렬식 (Average taxonomic distance와 Correlation coefficient 행렬식)을 이용한 dendrogram이 서로 다르다는 것이 확인되었다. 이들 두가지의 행렬식 중에서 taxa간의 형태적인 유연관계를 나타내는 하나의 dendrogram만을 선택하기 위한 한 객관적방법이 제안되었다. 즉 principal component analysis에 의한 결과를 비교할 표준결과로 이용하는 방법이다.

INTRODUCTION

Proctor (1966) stated that one of three main stages in a classification by numerical taxonomic processes is the calculation of a measure of resemblance between each pair of organisms, using the coded characters. Sneath and Sokal (1973) noted that the estimation of resemblances among taxa is the most important step in numerical taxonomy.

In cluster analyses to summarize similarities among populations, either average taxonomic distance matrix (Sokal, 1961) or correlation coefficient matrix was utilized. As the dendrograms constructed from these two matrices are different with each other, cophenetic correlation coefficient (Sokal and Rohlf, 1962) or similarity with the conventional taxonomy studied thus far (Ehrlich and Ehrlich, 1967) were used as a criterion to choose one of two matrices. These criteria are subjective, as stated by Sneath and Sokal (1973) that the choice among different similarity matrices is based on the worker's preference. An objective method to decide reliable similarity matrix for cluster analyses has been

suggested, using the result from principal component analysis as a standard.

MATERIALS AND METHODS

Materials:

571 adults of deer mice, *Peromycus maniculatus* (Cricetidae, Rodentia), selected from continental North America were used. Specimens from contiguous localities were arbitrarily grouped together and each specimen was assigned to one of 45 Operational Taxonomic Units, OTU's (for the localities of specimens collected, see Koh, 1980).

Characters:

Analyses were based on three external and 27 cranial characters including body length of tail vertebrae, the greatest length of the skull, and condylobasal length (for details refer to Koh, 1980). External measurements by collectors were used and cranial measurements were taken with a dial caliper to 0.01 millimeter.

Statistical analyses:

All computations were made using the University of Toronto IBM computer.

Sample statistices such as mean and standard error were computed by a high speed FORTRAN program (Koh, 1980). Clustering and ordination analyses were performed by using several subprograms of the Numerical Taxonomy System of Multivariate Statistical Program (NT-SYS) by Rohlf *et al.* (1974).

Raw data were first standardized using Sokal's (1961) equation, $(X_i-\bar{X})/\text{S.D.}$, where X_i indicates ith measurement, $\bar{X}$ the mean, and S.D. standard deviation (subprogram STAND of NT-SYS). Average taxonomic distance and correlation coefficient matrices were calculated from standardized data by using subprogram SIMINT of NT-SYS. The subprogram TAXON was used to group OTU's by using Unweighted Pair Group Method using Arithmatic averages, UPGMA (Sneath and Sokal, 1973). Principal component analysis (Seal, 1964) was performed using subprogram FACTOR of NT-SYS.

Cophenetic correlation coefficent between the original taxonomic distance matrix and the distance matrix from the dendrogram (Sokal, and Rohlf, 1962) or correlation coefficient between the average taxonomic distance matrix in original space and the distance matrix among OTU's in three dimensional projected space (Rohlf, 1972) were also calculated by subprogram MACOMP of NT-SYS.

RESULTS

Average taxonimc distance matrix calculated from standardized means of adults from 45 OTU's were used for UPGMA cluster analysis. The resultant dendrogram is shown in Fig. 1 (the cophenetic correlation coefficient was 0.87). The 16 OTU's (5, 15, 20, 23, 25,

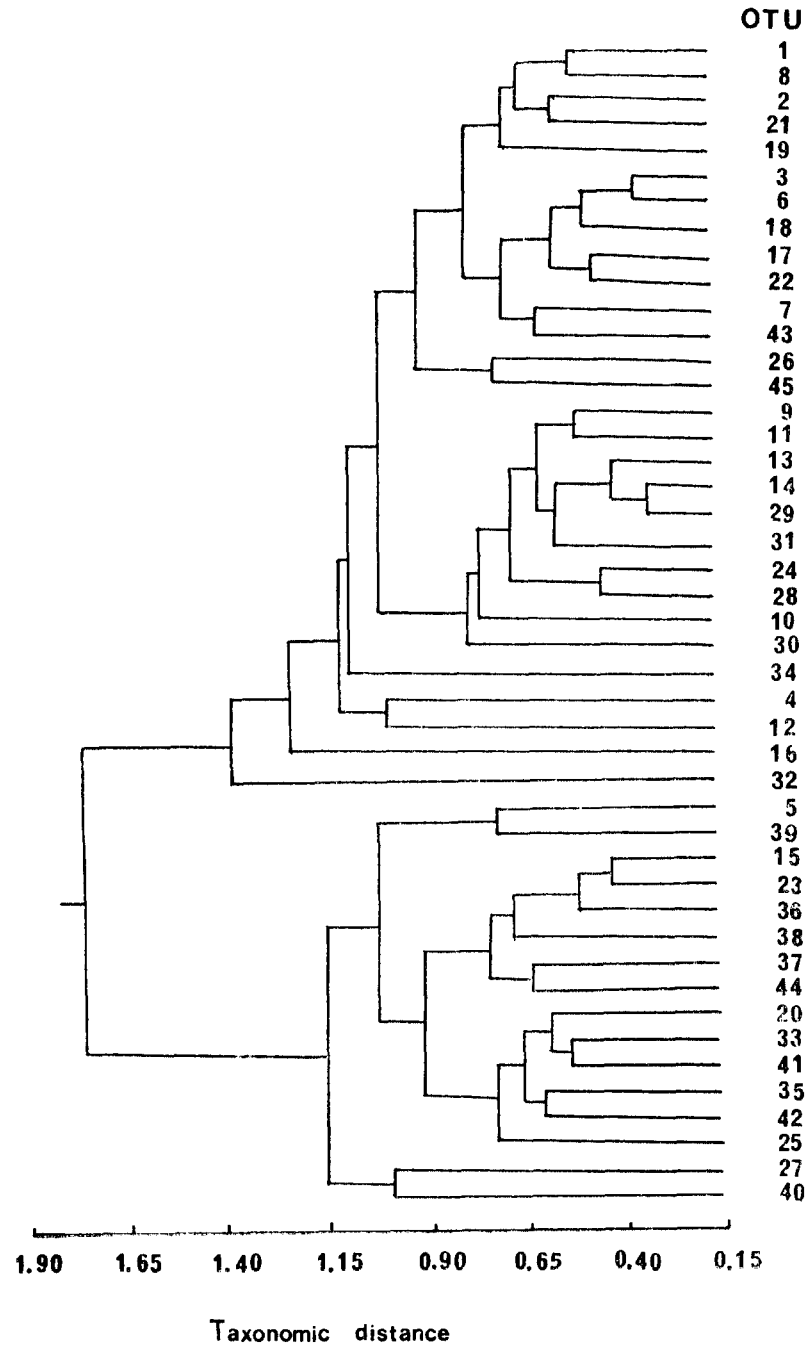

Fig. 1. Groupings of 45 OTU's of *P. maniculatus* based on UPGMA analysis using average taxonomic distance matrix from standardized means of adults. The cophenetic correlation coefficient was 0.87.

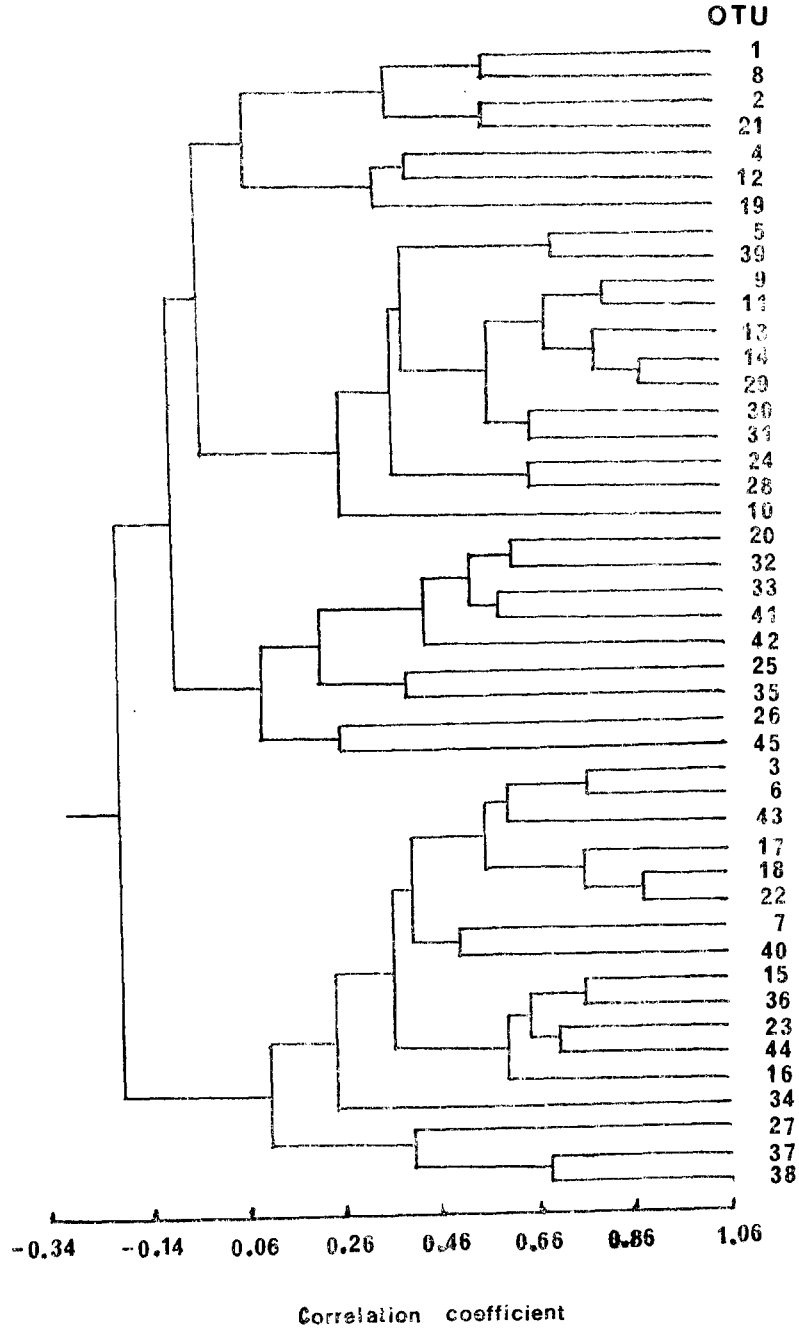

Fig. 2. Groupings of 45 OTU's of *P. maniculatus* based on UPGMA analysis using correlation coefficient matrix from standardized means of adults. The cophenetic correlation coefficient was 0.74.

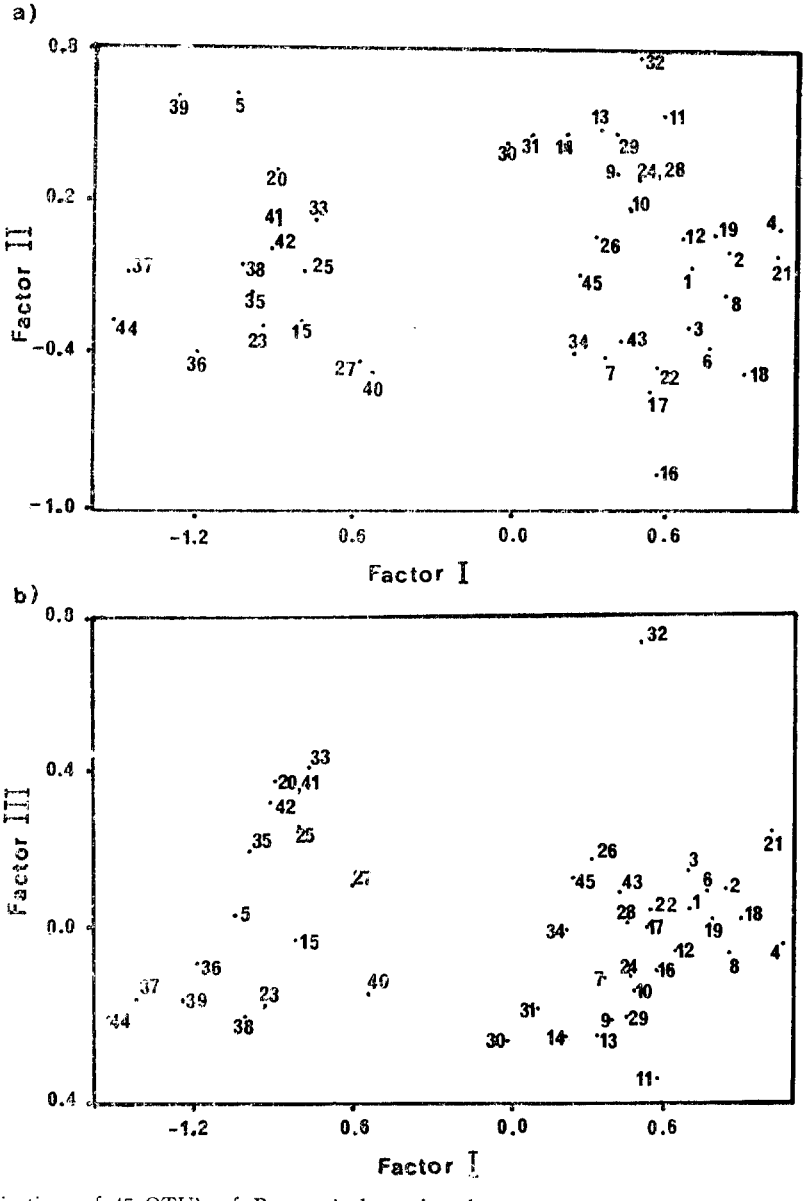

Fig. 3. Projections of 45 OTU's of *P. maniculatus* based on principal component analysis in three dimensions using standardized means of adults. Factors I, II, and III represented 60, 14, and 5 per cent of the variance, respectively. a) 45 OTU's ordinated with factor I vs. factor II. b) 45 OTU's ordinated with factor I vs. factor III.

27, 33, 35, 36, 37, 38, 39, 40, 41, 42, and 44) formed one group, and 29 OTU's (1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 16, 17, 18, 19, 21, 22, 24, 26, 28, 29, 30, 31, 32, 34, 43, and 45) formed another.

The correlation coefficient matrix obtained from the same data is shown in Fig. 2 (the

cophenetic correlation coefficient was 0.74). The OTU's composing two groups determined by the analysis using the average taxonomic distance (see Fig. 1) were sorted to form subgroups when the correlation coefficient was used for same analysis (see Fig. 2). The correlation coefficient between the average taxonomic distance matrix and the correlation coefficient matrix was -0.44.

Two-dimensional configuratioin from principal component analysis with the same data mentioned above is shown in Fig. 3 (factors I, II, and III represented 60, 14, and 5 per cent of the varience). The result from principal component analysis was the same as that from the UPGMA analysis using the average taxonomic distance, i.e., the OTU's composing two groups were the same. The correlation coefficient between the average taxonomic distance matrix in 30 dimensions and the distance matrix calculated from three-dimensional coordinates was 0.98.

In conclusion, the average taxonomic distance matrix rather than the correlation coefficient matrix is the matrix representing the similarity among taxa in cluster analyses using the data of deer mice.

DISCUSSION

As the basic tenets of numerical taxonomy include the use of resemblance (Sneath and Sokal, 1962), estimation of resemblance among taxa using either taxonomic distance matrix or correlation coefficient matrix in cluster analyses is the most important step. Moss (1967), however, found that dendrograms constructed from the same clustering method but different matrix, taxonomic distance or correlation coefficient, were different in acarin family Dermanyssidae. Lidicker (1973) concluded that numerical phenetic analyses must be used cautiously in view of their instability. Different groupings were produced, as shown in Figs. 1 and 2, with correlation coefficient of -0.44 between average taxonomic distance matrix and correlation coefficient matrix.

A criterion to choose one of similariy matrices which would represent phenetic relationship better than the other matrix is the value of cophenetic correlation coefficient between either the original taxonomic distance matrix or the original correlation coefficient matrix and the distance matrix from the dendrogram (Ehrich and Ehrlich, 1967). He choosed the dendrogram resulted from taxonomic distance matrix in butterflies because this dendrogram showed higher cophenetic correlation coefficent than the dendrogram from correlation coefficient matrix did. Farris (1969), however, stated that the cophenetic correlation coefficent should not be employed as an optimality criterion in classification.

Another criterion to determine one of similarity matrices is the degree of resemblance between the dendrogram constructed from one of similarity matrices and conventional taxonomic conclusions studied thus far (Cheetam, 1968). He stated that correlation coefficient matrix produced more realistic phenograms than distance matrix did in a

bryozoan genus *Mclrabdotus*. The criterion above can not be considered as one method to choose one of similarity matrices because numerical taxonomists are trying to show somewhat different phenetic relationship, which has been concealed under conventional taxonomy based on a few key characters. Moreover, Eades (1965) stated that the theoretical basis of the correlation coefficient matrix as a measure of resemblance is unsound, because there is a danger of fortuitous high coefficient when the taxa are not similar.

In short, there has been no known objective method to select one of two similarity matrices.

An objective method to determine one of two similarity matrices has been suggested; the groupings from these two matrices were compared with the result from principal component analysis as a standard result, and then the similarity matrix which is similar with the standard is considered as the one summarising phenetic relationship among taxa. It is found that two groups resulted from UPGMA analysis using the average taxonomic distance matrix were the same as those from principal component analysis (see Figs. 1, 2 and 3), with the correlation coefficient of 0.98 between the average taxonomic distance matrix in 30 dimensions and the distance matrix calculated from three-dimensional coordinates. It is concluded that the average taxonomic distance matrix rather than the correlation coefficient matrix is the matrix representing the similarities among taxa in phentic data of deer mice.

In further studies, the results from other ordination methods such as principal coordinate analysis and multidimensional scaling analysis will be compared with the result from cluster analyses. In addition, the groupings from different clustering methods have to be compared with one another, as noted by Oxnard (1978) that whatever the pattern constructed by a dendrogram, and however good the overall significance, the liability in the system make a number of other dendrograms possible.

SUMMARY

It has been confirmed that two dendrograms resulted from two similarity matrices, average taxonomic distance and correlation coefficient matrices, are different with each other when cluster analyses were performed with 571 adults of deer mice, *Peromyscus maniculatus* using 30 morphometric characters. To choose one of two similarity matrices mentioned above in order to construct a dendrogram representing phenetic relationships among taxa, an objective method using the result from principal component analysis as a standard result to compare with two matrices has been suggested.

REFERENCES

Cheetam, A.H., 1968. Morphology and systematics of the bryozoan genus *Metrarabdotos*. *Smithonian Misc. Coll.*, 153(1). pp. 121

Eades, D.C., 1965. The inappropriateness of the correlation coefficient as a measure of taxonomic resemblance. *Syst. Zool.*, **14**:98-100.

Ehrlich, P.R. and A.H. Ehrlich, 1967. The phenetic relationships of butterflies. I. Adult taxonomy and the nonspecificity hypothesis. *Syst. Zool.*, **16**:301-317.

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

Koh, H.S., 1980. A phenetic study of deer mice, *Peromyscus maniculatus* Wagner (Cricetidae, Rodentia), from continental North America, with chromosomal analyses from four Canadian populations. Ph. D. disseration, Univ. of Toronto, Canada.

Lidicker, W.Z., 1973. A phentic analysis of some new guinea rodents. *Syst. Zool.*, **22**:36-45.

Moss, W.W., 1967. Some new analytic and graphic approaches to numerical taxonomy, with an example from the Dermanyssidae (Acari). *Syst. Zool.* **16**:177-207.

Oxnard, C.E., 1978. One biologist's view of morphometrics. *Ann. Rev. Ecol. Syst.*, **9**:219-241.

Proctor, J.R., 1966. Some processes of numerical taxonomy in terms of distance. *Syst. Zool.*, **15**: 131-140.

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

Rohlf, F.J., J. Kishpough, and D. Kirk., 1974. Numerical taxonomy system of multivariate statistical programs. Stoney Brook, New York.

Seal, H. L., 1964. Multivariate statistical analysis for biologists. J. Wiley, New York.

Sneath, P.H.A. and R.R. Soakl., 1962. Numerical taxonomy. *Nature*, **193**:855-860.

Sneath, P.H.A. and R.R. Soakl., 1973. Numerical Taxonomy. W. H. Freeman and Co., San Francisco.

Sokal, R.R., 1961. Distance as a measure of a taxonomic similarity. *Syst. Zool.*, **22**:360-374.

Sokal, R.R., and F.J. Rohlf., 1962. The comparison, of dendrograms by objective methods. *Taxon* **11**:33-40.