THE EFFECT OF USING VARIOUS CHARACTER SUBSETS ON NUMERICAL TAXONOMY IN THE JAPANESE ANDRENID BEES

Tadauchi, Osamu

THE EFFECT OF USING VARIOUS CHARACTER SUBSETS ON NUMERICAL TAXONOMY IN THE JAPANESE ANDRENID BEES

OSAMU TADAUCHI

Entomological Laboratory, Faculty of Agriculture,
Kyushu University, Fukuoka 812, Japan

Abstract

The effect of using eleven subsets of characters on the clustering method was tested based on 85 taxa or OTUs of the Japanese Andrenid bees (Hymenoptera, Andrenidae). The eleven character subsets, which consisted of total (Nos.=130), randomly selected (100, 70, and 40), head-thoracic (40 and 58), total structural (40 and 82), PCA produced (21-weighted and 21-unweighted) and key characters (21), were analyzed by cluster analysis based on distance coefficients. The results were evaluated and discussed at the subgeneric level of the genus *Andrena*. The two phenograms derived from the 100 and 130 character subsets were similar to each other. The two PCA produced subsets not necessarily resulted in enough groupings, The 21 key character subset produced different clusters far from that derived from the total characters.

Introduction

Sneath and Sokal (1973) expressed that they could not justify the recommendation of use of no less than 60 characters (presented by Michener and Sokal, 1957) on either empirical or theoretical grounds and that they were unable to provide generally valid answers to the question of requisite number of characters.

The purpose of the present study is to examine the effect of using different character subsets on the clustering method as applied to Japanese Andrenid bees. Furthermore, the present analyses are designed as a test of using principal component scores produced by principal component analysis (PCA).

Material and Method

Material and Characters

The original data used in the present study were obtained as a by-product of numerical taxonomy in the genus *Andrena* of Japan (Tadauchi, 1982). One hundred and thirty characters used in the present study are commonly employed for the taxonomy in the genus *Andrena*. Eighty-five taxa or OTUs and the 130 characters

---

used were listed in the previous paper (Tadauchi, 1982, Tables l-2). The eleven subsets of characters are as follows :

A. Total characters
    1. 130 original subset : original characters (Tadauchi, 1982, Table 2, Code Nos. 1– 130).

B. Randomly selected characters
    2. 100 random subset : 100 characters randomly selected from the original data using random numbers.
    3. 70 random subset : 70 random characters selected as above.
    4. 40 random subset : 40 random characters selected as above.

C. Head-thoracic characters
    5. 40 head subset : 40 characters (33 structural and 7 pubescence) derived from head region only (Tadauchi, 1982, Table 2, Code Nos. 2-34 & 83-89).
    6. 58 thoracic subset : 58 characters (35 structural and 23 pubescence) derived from the thoracic region only (Tadauchi, 1982, Table 2, Code Nos. 35-69 & 90-111).

D. Total structural characters
    7. 40 hair subset : 40 characters related to pubescence on the body (Tadauchi, 1982, Table 2, Code Nos. 83-122).
    8. 82 structural subset : 82 characters (one body sized and 81 structural) related to structures of the body (Tadauchi, 1982, Table 2, Code Nos. l-82).

E. PCA produced characters
    This data group is prepared for a test of reduction of the number of characters. From PCA in my previous work (Tadauchi, 1983), 21 principal components were obtained accounting for 80.32 % of the total variance among the 130 original characters. Therefore the 130 basic characters may be able to reduce to 21 data or new characters in spite of excluding about 20 % of the total variance. For this purpose the following two subsets are prepared.
    9. 21 PC score subset (weighted) : 21 principal component scores of OTUs for the 21 PCs, weighting with percentage value corresponding with each PC.
    10. 21 PC score subset (unweighted) : the same data as above, weighting equally.

F. Key characters
    11. 21 key character subset : 21 key characters usually used for the subgeneric level of the genus *Andrena.*

Method

    Each of the eleven subsets concerning to the 85 OTUs was summarized by a group average method of cluster analysis in SAC (Tadauchi, 1981), using standardized Euclidian distance and weighted Euclidian distance coefficients. The resulting distance phenogram derived from the 130 original subset was used as a standard to compare with the phenograms based on the other character subsets.

    All computations were carried out on FACOM M-200 computer at the Computer Center of Kyushu University.

Results and Discussion

Figure 1 and Figs. 2 to 11 show the 11 distance phenograms. Figure 1 is derived from the 130 original subset. The detail result was explained in my previous paper (Tadauchi, 1982). A line drawn through the phenogram corresponding to a distance value of 1.25 divided the OTUs into groupings identical with most of the recognized subgenera. Figure 2 represents the phenogram obtained from the 100 random subset. The results were considerably similar to the preceding phenogram. The OTUs of the subgenera, *Simandrena* (group code SIM in Figs. and Table), *Euandrena* (EU), *Hoplandrena* (HOP), **Cnemidandrena** (CNE), *Micrandrena* (MIC) and others were tightly together with one another. The difference was the strong separation of the **mikado A-ishihurai (2-4)** cluster from the other cluster of the same subgenus. In the **Gymnandrena** (GYM), *parathoracica (30)* was relatively isolated from the other OTUs of the same subgenus. The OTU **fukuokensis (55)** was more closely related to the *nitidiuscula-richardsi (50-51)* cluster. The result obtained from the 70 random subset is shown in Fig. 3. This phenogram basically agreed with the previous two analyses. The relationships among the OTUs at the subgeneric level remained invariant for the most part. The important differences evident in this phenogram were as follows :1) the OTUs of the subgenus *Andrena* (AND) clustered tightly together, 2) *Andrena* sp. (69) isolated from the *Larandrena* (LAR, 63-67) cluster and joined the *Andrena* cluster,

Table 1. Groupings obtained by different character subsets. The presence of the same group as that from the 130 original subset at a 1.25 distance level is denoted by a "＋", and its absence by a "－".

| Group Code	130 Orig.	100 Rand.	70 Rand.	40 Rand.	40 Head	58 Thor.	40 Hair	82 Str.	21 PC-w.	21 PC-uw.	21 Key
AND	+	−	+	+			−				
LAR	+	−									
EU	+	+	−				+		+	+	
HOP	+	+					−		+		
CNE	+	+	+	+	−	+	+	+	+	+	+
GYM	+	−	−		−			+	+		
SIM	+	+	+	+	+			+	+	+	
MIC	+	+	+	+	−			+	+	+	
NOT I	+			−						+	
POE	+				−				−		
NOT II	+	+		−	−	−	−	+	−	+	
CAL	+		+			+					−
ORE	+		+					+			
TAE	+	+	+	+	+			+			
HAB	+	+	+	+	−	+	−	+	+	+	+
CHL	+	+			+			+	+	+	+
STE	+	+	+		+		+	+	+	+	−
MIT	+	+	+	+	+			+	+	+	+
PLA	+	+	+	+	−	−		+	+	+	+
TRA	+	+	+	+		+		+	+	+	+
HOL I	+	+	+	−		+				+	
HOL II	+	+	+					−	−	+	
PAR	+	+	+	+	+	+	+	+	+	+	+
Total	23	16	15	10	6	6	4	14	13	15	7

3) tateyamana (Euandrena sp. 1, in my previous paper, Tadauchi, 1982) joined the *Larandrena* cluster before connecting with *Euandrena (24-73)* cluster. The phenogram based on the 40 random subset is shown in Fig. 4. Through an examination of the phenogram, many differences could be recognized in comparison with the preceding three results. One main difference was transference of OTUs at the subgeneric level, for instance, for *sasakii* (33) joining the *Hoplandrena (38-76)* cluster, and for *tateyamana (70)* clustering with *Chlorandrena* (CHL, 18-20) cluster. The other main difference was closer relationships among groups throughout the phenogram, for instance, for *Calomelissa* (CAL, 14-15) cluster connecting with *Oreomelissa* (ORE, 16-62) cluster at a relatively low distance value of 1.035, and for *halictoides (58)* joining the *Larandrena (63-67)* cluster at a low distance value of 1.152. Although many differences were seen, rough grouping of the major clusters at the subgeneric level was still maintained in this phenogram.

To compare the above three phenograms with that derived from the 130 original subset, the phenon line is also used drawing at a distance level of 1.25. Table 1 shows the groupings obtained by the eleven different character subsets. In the table the presence of the same group constructed as that from the *130* original subset at the 1.25 level is denoted by a " $+$ ", and its absence by a " $-$ ". As shown in the table, 23 groups were produced from the basic subset. While 16, 15, and 10 groups were obtained from 100, 70, and 40 random subsets, respectively. In the present study it seems that a change of character numbers produces different clusterings as mentioned by Crovello (1969) and Tadauchi (1978). However, a closer observation of the table from the right (the 40 random subset) to the left (the 130 original subset) shows that there is one direction in formation of groupings except for a few subgenera *(Andrena, Calomelissa,* and *Oreomelissa).* Namely, the more the characters are employed, the closer the groups produced are to those from the 130 original subset. Nine subgenera are already obtained from the 40 random subset as follows : *Cnemidandrena, Simandrena, Micrandrena, Taeniandrena* (TAE), *Habromelissa* (HAB), *Mitsukuriella* (MIT), *Plastandrena* (PLA), *Trachandrena* (TRA), and *Parandrena* (PAR). Next, three subgenera or groups, *Stenomelissa* (STE), *valeriana (35,* HOL I), and *ishikawai-tuniguchiae (36-37,* HOL II) cluster are formed from the 70 random subset. Four subgenera or groups, i. e., *Euandrena, Hoplandrena, amamiensis (52, NOT* II), and *Chlorandrena* are estab-Iished from the 100 random subset. Finally *Larandrena, Gymnandrena, richardsi (51,* NOT I) cluster and *Poecilandrena* (POE) are produced derived from the 130 original subset. As shown in the previous paper (Tadauchi, 1982), the last four subgenera or groups are changeable even by the same data using by different clustering methods. If the phenon line is lowered from *1.25* to 1.15, 21 out of 23 groups derived from the original subset are produced from the 100 random subset. So the present study shows that relatively stable groupings are obtained from the 100 characters. The 130 original subset had 21 principal components (PCs) at a level of 80.00 accumulated percentage (Tadauchi, 1983). The 100, 70, and 40 random subsets had 20, 17, and 14 PCs at the same level, respectively. Taxonomic information content (considered from the numbers of PCs and the character groups gathered by PCA) of the 100 random subset was very similar
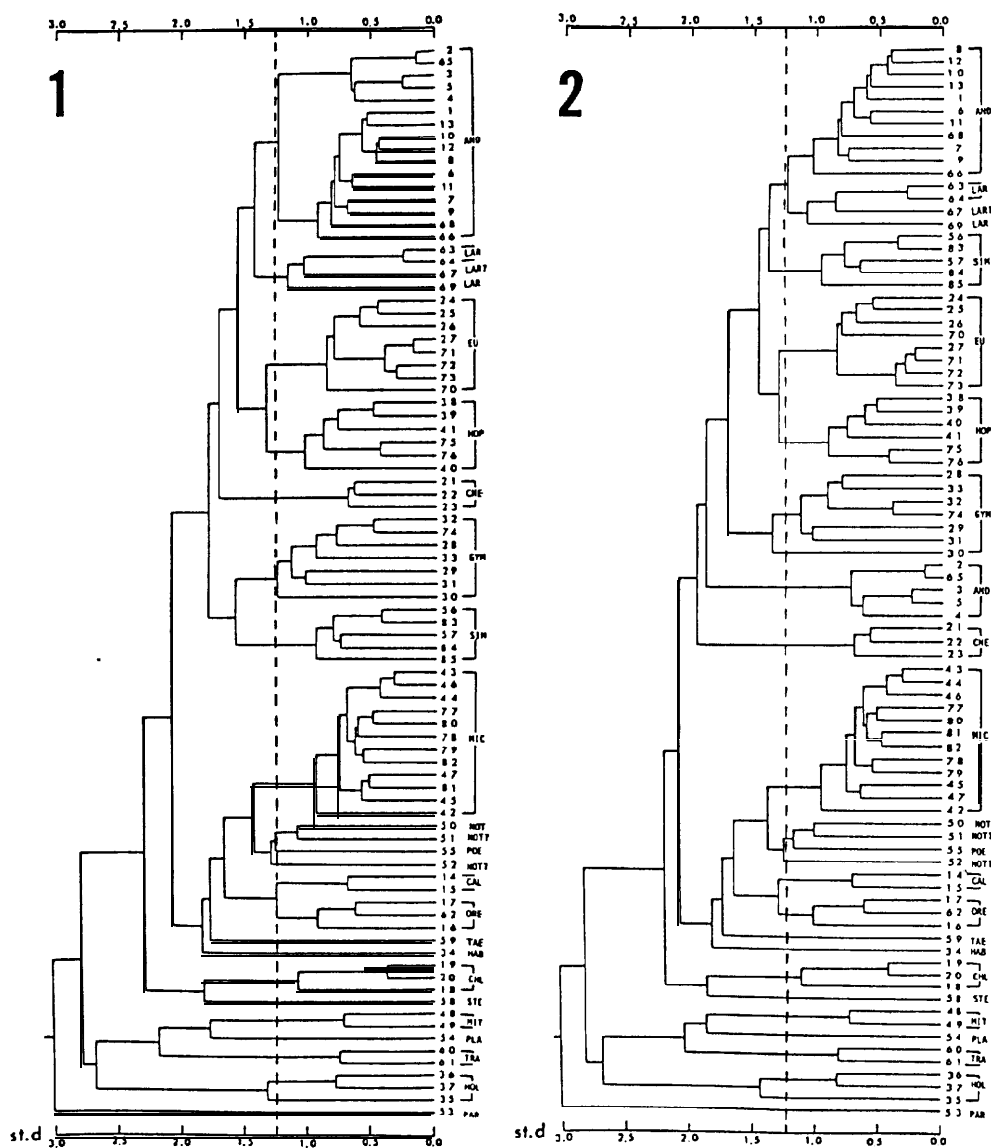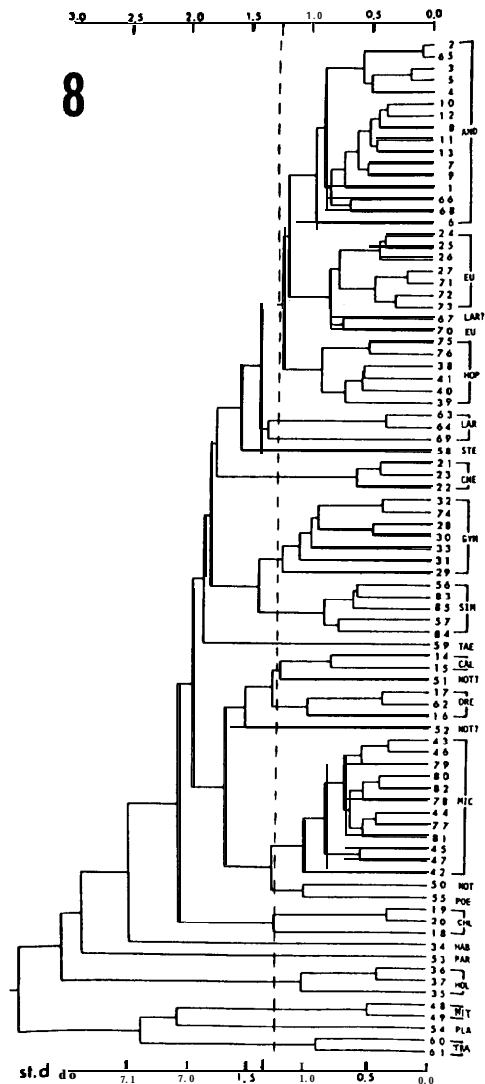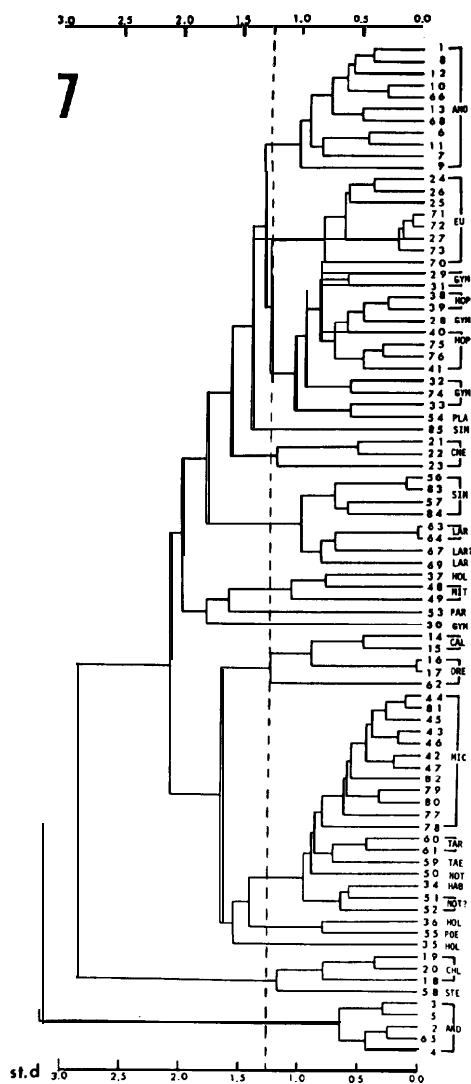
Figs. 1-2. Distance phenograms obtained from 130 original (1) and 100 random (2) subsets derived from 85 OTUs of Japanese Andrenid bees, using the group average method of cluster analysis. A dashed line shows comparing distance level (*st. d.* = 1.25).

Figs. 3-4. Distance phenograms obtained from 70 random (3) and 40 random (4) subsets derived from 85 OTUs of Japanese Andrenid bees, using the group average method of cluster analysis. A dashed line shows comparing distance level (*st.d. = 1.25*).
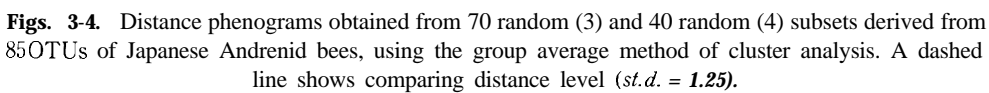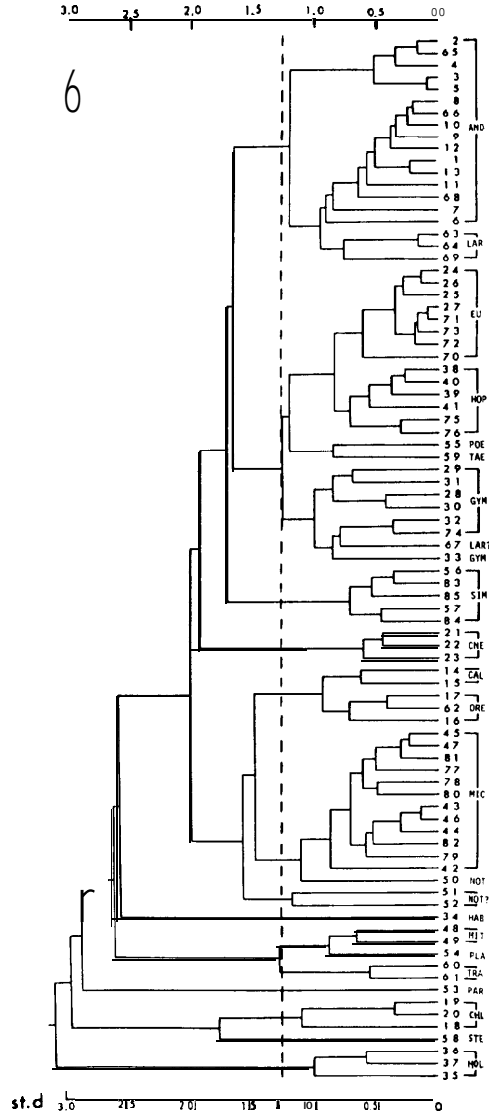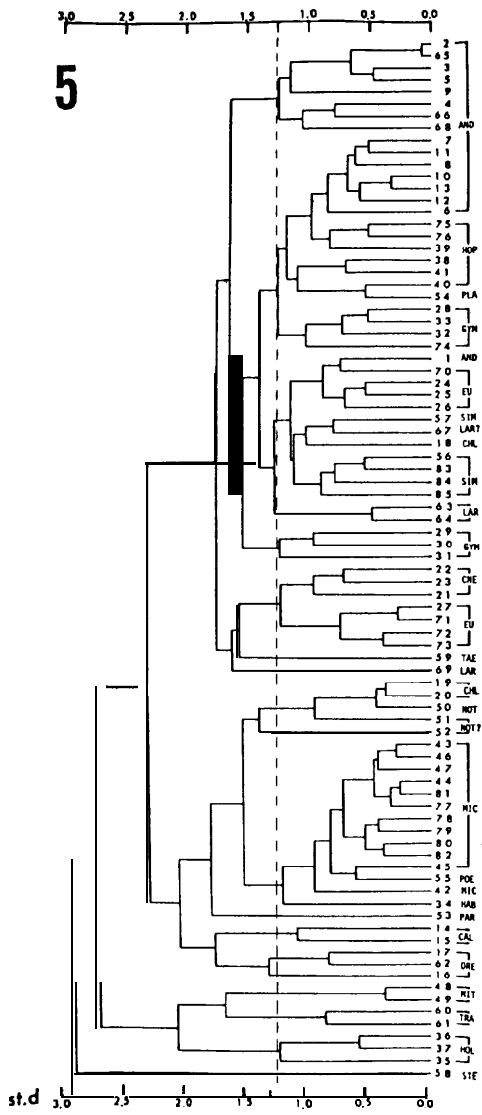
Figs. 5-6. Distance phenograms obtained from 40 head (5) and 58 thoracic (6) subsets derived from 85 OTUs of Japanese Andrenid bees, using the group average method of cluster analysis. A dashed line shows comparing distance level (st. $d.$ = 1.25).

Figs. 7-8. Distance phenograms obtained from 40 hair (7) and 82 structural (8) subsets derived from 85 OTUs of Japanese Andrenid bees, using the group average method of cluster analysis. A dashed line shows comparing distance level (*st. d.* = 1.25).

Figs. 9-10. Distance phenograms obtained from **21** PC score subsets, weighted (9) and unweighted (10), derived from 85 OTUs of Japanese Andrenid bees, using the group average method of cluster analysis. A dashed line shows comparing distance level *(weighted d. = 20.0* for Fig. 9, and 4.0 for Fig. 10).
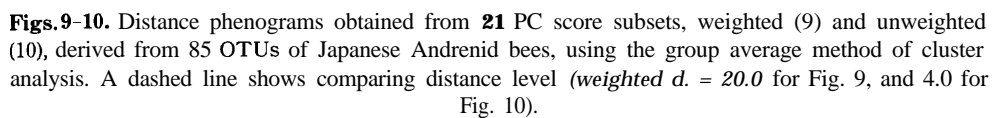
Fig. II. Distance phenogram obtained from 21 key character subset derived from 85 OTUs of Japanese Andrenid bees, using the group average method of cluster analysis. A dashed line shows comparing distance level (st. $d. = 1.25$).

to that of the 130 original subset. Therefore it is reasonable that the results obtained from the 100 and the 130 character subsets are similar to each other. The 70 random subset had fewer information content than the previous two subsets. The 40 characters had the fewest information content in the four subsets. According to the earlier hypothesis of the matches asymptote presented by Sneath and Sokal (1962), the value of the similarity coefficient becomes more stable as the number of characters sampled increases. The present study shows that at least 100 morphological characters may be necessary to produce groupings at the subgeneric level of *Andrena*. An acceptable classification should possess a stability when new information are introduced. It seems necessary to study the other kinds of characters, such as genetical, ecological (floral, seasonal, etc.) and so on.

Figures 5 and 6 show the distance phenograms obtained from the 40 head and the 58 thoracic subsets, respectively. At the phenon line of 1.25, 18 groups were produced from the 40 head subset. However, only six groups, which were the same as those based on the 130 original subsets, were obtained. There were many differences as for the formation of groupings in comparison with those by the 130 characters. Many subgenera were divided into two or three separate groups. For instance, the subgenus *Andrena* was divided into three clusters (2-68, 7-6, and 1 clusters). The phenogram from the 58 thoracic subset also had six groups identical with those derived from the 130 original subset. Rough grouping of the major clusters at the subgeneric level was maintained in this phenogram.

Figures 7 and 8 show the distance phenogram based on the 40 characters of hairs and the 82 characters of structures. The phenogram from the 40 hair subset showed many displacements of OTUs. For instance, the subgenus **Gymnandrena** was divided into three separate groups. The same groups identical with those from the 130 original subset were only four. On the other hand, the phenogram from the 82 structural subset had 14 groups which were the same as those obtained by the 130 characters. This results were similar to those obtained by the 70 random subset. By the test of the hypothesis of nonspecificity (e.g., Michener and Sokal, 1966, Ehrlich and Ehrlich, 1967), it seems to be obvious that different character sets chosen from different regions in the body will give somewhat different relationships. In the present study three different subsets consisted of 40 characters, i. e., 40 random, 40 head, and 40 characters of hairs, respectively, produced different groupings with one another. It seems for me that the amount of 40 characters are not enough to obtain the groupings at the subgeneric level as mentioned above.

Figure 9 shows the distance phenogram based on the 21 PC score subset, weighting with percentage corresponding to each PC. As the phenogram obtained by using weighted distance coefficient, it can not be equally compared with the phenogram from the original subset. So if a weighted distance line of 20.0 is cited as a phenon line in Fig. 9, 13 out of 23 groups are identical with those obtained from the 130 subset. This results are similar to those derived from the 70 random subset. Figure 10 shows the phenogram from the same data as above, weighting equally. If the weighted distance line of 4.0 is cited in this case, it produces 15 groups identical with those from the 130

characters. This result is also similar to that based on the 70 random subset. The 21 PC scores include about 80 ％ among the character variance of the 130 total characters. These results are not necessarily satisfactory for the subgeneric classification of **Andrew.** It shows that the remaining 20 ％ of character variance can not be ignored. Finally Fig. 11 shows the phenogram from 21 key character subset, to compare with the above 21 PCs results. Only seven groups identical with those from the 130 original subset were obtained. Although the 21 key characters used were generally considered important for the taxonomy of *Andrena,* the result is different far from that based on the 130 total characters and current classification. It seems that the number of characters are fewer and that may be necessary to use more characters.

Acknowledgements

References

Crovello, T. J., 1969. Effects of change of characters and number of characters in numerical taxonomy. *Amer. Midland Nat., 81: 68-86.*

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

Michener, C. D. and R. R. Sokal, 1957. A quantitative approach to a problem in classification. *Evolution,* 11: 130-162.

Moss, W. W., 1968. Experiments with various techniques of numerical taxonomy. *Syst. Zool.,* 17 : 31-47.

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

———— and ————, 1973. Numerical *Taxonomy.* W. H. Freeman and Company, San Francisco.

Tadauchi, O., 1978. The effect of using different numbers of characters on the clustering method as applied to Japanese Andrenid bees. *Zool. Mag.,* 87 : 186-190.

———— 1981. Taxonomic working system by computer (SAC) with application to Japanese Andrknid bees. *Esakia, (17) : 161-182.*

————, 1982. A numerical taxonomic study of the genus *Andrena* (Hymenoptera, Andrenidae) of Japan. *J. Fac. Agr., Kyushu Univ.,* 26: 169-191.

————, 1983. Summarization of taxonomic information in the Japanese Andrenid bees by principal component analysis. *Kontyû, Tokyo, 51: 351-157.*