



<http://www.biodiversitylibrary.org>

**The University of Kansas science bulletin.**

[Lawrence] :University of Kansas,1902-1996.

<http://www.biodiversitylibrary.org/bibliography/3179>

**38, pt. 2:** <http://www.biodiversitylibrary.org/item/23745>

Page(s): Page 1409, Page 1410, Page 1411, Page 1412, Page 1413, Page 1414, Page 1415, Page 1416, Page 1417, Page 1418, Page 1419, Page 1420, Page 1421, Page 1422, Page 1423, Page 1424, Page 1425, Page 1426, Page 1427, Page 1428, Page 1429, Page 1430, Page 1431, Page 1432, Page 1433, Page 1434, Page 1435, Page 1436, Page 1437, Page 1438

Contributed by: Harvard University, MCZ, Ernst Mayr Library

Sponsored by: Harvard University, Museum of Comparative Zoology, Ernst Mayr Library

Generated 9 February 2009 12:50 PM

<http://www.biodiversitylibrary.org/pdf1/000194800023745>

This page intentionally left blank.



# THE UNIVERSITY OF KANSAS SCIENCE BULLETIN

VOL. XXXVIII, Pt. II] MARCH 20, 1958

[No. 22

---

---

## A Statistical Method for Evaluating Systematic Relationships<sup>1</sup>

BY

ROBERT R. SOKAL and CHARLES D. MICHENER<sup>2</sup>

Department of Entomology  
University of Kansas, Lawrence

**ABSTRACT.** Starting with correlation coefficients (based on numerous characters) among species of a systematic unit, the authors developed a method for grouping species, and regrouping the resultant assemblages, to form a classificatory hierarchy most easily expressed as a treelike diagram of relationships. The details of the method are described, using as an example a group of bees. The resulting classification was similar to that previously established by classical systematic methods, although some taxonomic changes were made in view of the new light thrown on relationships. The method is time consuming, although practical in isolated cases, with punched-card machines such as were used; it becomes generally practical with increasingly widely available digital computers.

### INTRODUCTION

The purpose of the study reported here was to develop a quantitative index of relationship between any two species of a higher systematic unit, as well as to exploit such indices of association in the establishment of a satisfactory hierarchy. The authors became interested in the development of such a method when they attempted to find a technique for classifying organisms that was free from the subjectivity inherent in customary taxonomic procedure.

---

1. Contribution number 945 from the Department of Entomology, University of Kansas.

2. We wish to acknowledge the constructive criticism received in connection with this and related work from the following individuals who kindly gave their time to read and comment upon the manuscript: Paul R. Ehrlich, University of Kansas; Raymond B. Cattell, University of Illinois; Alfred E. Emerson, University of Chicago; Warwick E. Kerr, Universidade de São Paulo; Ernst Mayr, Harvard University; Louis L. McQuitty, Michigan State University; G. G. Simpson, American Museum of Natural History; Peter C. Silvester-Bradley, University of Kansas and University of Sheffield; and Paulo E. Vanzolini, Departamento de Zoologia, Secretaria de Agricultura, São Paulo. These persons, however, are not responsible for the opinions which we have expressed.

Acknowledgment is also due to the University of Kansas General Research Fund for assistance.



The systematic group chosen as a test of the feasibility of this undertaking was one consisting of 97 species of solitary bees in the family Megachilidae. This choice was made because one of us (C. D. M.) has made recent systematic studies of these insects, so that conclusions as to the relationships obtained by the usual systematic procedure could be compared with the results of the new method.

The findings of our study as well as the philosophical bases of our attempts at quantifying systematic relationships have been reported elsewhere (Michener and Sokal, 1957). In this paper we propose to describe in some detail the actual method employed, as well as our reasons for adopting it and for rejecting several alternate procedures. It is our intention to illustrate the procedures in sufficient detail so that persons with a limited knowledge of statistical methods will be able to follow our method. We expect our system to be applicable to most organisms, provided they exhibit a variety of characters, and the account to follow is consequently phrased in general terms. However, our practical illustrations are based on the bee group cited above in order to provide the reader with concrete examples.

A quantitative method of finding the relationship between two species must be based on a number of taxonomic characters in a manner similar to the traditional systematic approach. However, whereas the latter technique generally uses few characters and weights these quite unequally and subjectively, the former method employs numerous but unweighted characters. Our reasons for not weighting characters have been detailed in the companion paper (Michener and Sokal, 1957). In the absence of an objective criterion of character weight it seems best to rely on a large number of equally weighted characters. In our bee study we employed 122 characters per species; however, we feel significant results may be obtained from as few as 60 characters.

Our use of the word "character" will require some elaboration. In its commonest taxonomic usage, a character is any feature of one kind of organism that differentiates it from another kind. Thus the red abdomen of one bee is a character distinguishing it from another bee with the abdomen black. In this paper we use the word in a second connotation only; that is, as a feature which varies from one kind of organism to another. Now, to use the above example, abdominal color is the character, which occurs in two "states" or alternatives, red and black.



For each character the states were coded: 1, 2, 3, etc. In the bee study the number of states per character ranged from two to eight. Much variation in the number of states is undesirable from the point of view of the methods discussed below. In the study we undertook most characters had either three or four states. However, when variation exceeds desirable bounds it might be preferable to divide the character state codes by a common denominator or to normalize them.

The kinds of characters used in the bee study and the manner in which they were coded are discussed at length by Michener and Sokal (1957). The possible effect of parallelism is also treated in the same article. For purposes of the present paper the available data might be summarized as follows: we have records of a given number ( $n$ ) of species. For each species we have  $k$  records,  $k$  being the number of characters considered in the study. The coded values for any character may range from 1 to 9 depending on the number of states in which this character occurs in the group under consideration. As was mentioned previously it is desirable to have the number of states not differ too widely for the various characters. While it is not necessary to limit the number of possible character states to nine, our particular computational setup was greatly facilitated by the use of a single digit code.

## PROCEDURES

### *Character correlations and species correlations*

Two obvious ways suggested themselves to the authors regarding a procedure for deducing relationships from the character states of a group of species. We could either correlate characters with each other or species with each other. Since both of these methods would lead to interpretable, although differing, results a brief discussion of the implications of the two approaches follows.

Sturtevant (1942) undertook a study of the genus *Drosophila* with objectives and procedures somewhat similar to ours. He recorded 33 morphological, cytological and life history characters for each of 56 species of *Drosophila* and two species of the genus *Scaptomyza*. In his aim to develop a classification "as free from personal bias as I could make it," Sturtevant set up two tables. The first was a table of the total number of differences with respect to the 33 chosen characters between any two of the 58 species. These give the degree of difference between the species concerned and are



analogous to the complementary values of the "matching coefficient" discussed in the section on Choice of a Correlation Coefficient below.

A second table showed correlations between characters, expressed as two-way frequency distributions. By examining the three highest character correlations Sturtevant found that six species consistently fell into the exceptional classes of the two-way frequency distributions. They were the two *Scaptomyza* species and four species of *Drosophila* which he thereupon placed in separate subgenera. On the basis of the number of character differences between and within subgenera Sturtevant was able to confirm this classification and arrive at some ideas on the relationships and origins of the various groups. He also performed a similar analysis on 29 characters of 40 genera of flies (*Scatophaga*, *Conops* and 38 assorted Acalypterae) to establish the relations of the family Drosophilidae. Unfortunately the paper cited lists only summaries of the above tables and it is therefore difficult to compare Sturtevant's findings with ours.

Correlation between characters (R-technique in the idiom of the factor analysts) is the customary technique in biological and psychological studies involving correlational analysis. In character correlation matrices involving studies within one species each correlation represents the sum total of the common forces acting on any pair of characters. When analyzed by some method of factor analysis, the matrix customarily yields a so-called general size factor, a series of group factors affecting various groups of characters, and residual specific factors affecting single characters only. The foregoing is an example of a factor constellation involving morphological characters and is not necessarily the only possible constellation. As a matter of fact much psychometric work and the biometric papers by Howells (1951) and Stroud (1953) use the method of "simple structure" which *a priori* rejects solutions involving general factors.

Regardless of the constellation preferred, the factors common to two characters and causing them to be correlated could be visualized as developmental forces, genetic or environmental in the final analysis. The range of these genetic or environmental forces is dependent on the causes of variation within the sample of individuals studied. Thus a sample of individuals from an inbred, isogenic, line of animals would yield character correlations reflecting common nongenetic, physiological (*i. e.*, caused by microecological dif-



ferences) factors only. Another sample comprising individuals from various races or subspecies would provide correlations based on common factors representing (1) genetic differences between individuals; (2) genetic differences between races; (3) nongenetic physiological differences between individuals and (4) nongenetic ecological differences between races. One of the authors (R. R. S.) has been able to accumulate a series of character correlation matrices from various organisms representing these levels of variation. Matrices on correlation of aphid characters within galls (clones) and between galls have been published (Sokal, 1952) while similar matrices on aphid correlations between localities and morphological correlations within and between strains of houseflies and *Drosophila* await suitable analysis and publication.

When the sample transcends the bounds of the species the factors behind a character correlation matrix take on new meaning: They now represent genetic divergence or the results of evolutionary processes. In the one case they were ontogenetic forces, in the other they are phylogenetic forces. This type of analysis was pioneered by Stroud (1953) who analyzed correlations of 14 characters for soldiers of 48 species and imagines of 43 species of the termite genus *Kaloterme*s. He was able to interpret some factors extracted from his correlation matrices as recognizable evolutionary trends.

Another method of correlational analysis is called the transposed matrix method or the Q-technique (as compared with the R-technique of character correlations, discussed above).<sup>3</sup> It consists of correlations between individuals based on measurements of characters which they have in common. In psychology this involves correlations between persons based on scores for common tests which these persons have taken. In the Q-technique we are in effect dealing with the same kind of raw data as in the R-technique, but we compute the correlation coefficients by summing squares and products at right angles to the direction previously taken (or we transpose the matrix before computation which amounts to the same thing).

A Q-technique correlation coefficient in a study correlating individuals of one species represents common forces or factors acting on the two individuals concerned. In this case we cannot speak of the "sum total of common forces" as we could in the case of the

---

3. In a recent paper Cattell (1954) has suggested restricting the Q and R symbolism to studies involving factor analysis and proposed Q' and R' for studies, such as the present one, employing more superficial methods.



R-technique. Insofar as the characters used are indicative of the entire spectrum of potential variation of the individuals we can say that the resulting correlation coefficient is representative of the real affinity between two individuals. When scanned for clusters of high correlation coefficients the Q-type matrix reveals types of individuals which are similar. It is thus especially suited to classificatory problems. When subjected to factor analysis the resulting factors are now of a different nature. The general size factor has been lost and in its place we find a general taxonomic group factor which accounts for the overall correlations of all the individuals in the study.

When, as in the present study, the correlation is between species of a taxonomic unit the general factor is a general systematic factor denoting overall relationship within the systematic group. The species having the highest factor loading would be most representative of the group. Other factors would describe subgroups within the systematic unit and describe the relationships of these subgroups with each other and of the species to the subgroups. It should be clear from the above that for purposes of biological classification the relationships represented by a Q-technique matrix are more meaningful by far than are those of a R-technique matrix. Except for the above-mentioned work of Sturtevant (1942) which involved not correlations but character differences, the only Q-type study in systematics of which the authors are aware is in a publication by one of them (Sokal, 1958) containing factor analyses of selected portions of the present data. A number of the phytosociological coefficients of association and similarity can be considered as of the Q-type.

Psychologists have used Q-technique repeatedly (*e. g.*, Burt 1937, Stephenson 1936), although R-technique is still preferred in most studies. Cattell (1952) has listed 5 points of criticism of the Q-technique. It is appropriate that we discuss briefly their relation to the problems under study here. The first objection is that Q-technique loses the general size factor, yielding in its place a common species factor. This latter is claimed to be trivial by Cattell, and correctly so, for psychological work. However, in a matrix of correlations between species such a general systematic factor delineates the relation of individual species to the taxonomic group and indicates the proportion of the variance of each species explained by the general systematic factor.

Cattell's second objection to Q-technique is that it is unreasonable to assume simple structure in the factorization of a Q-matrix. The



authors agree with this argument, but for the purposes of the present paper it is not important since they are not here undertaking a factor analysis. Furthermore, they feel that simple structure (*i. e.*, each factor having essentially zero effects on certain species) is not necessarily a very suitable constellation for many biological factorizations.

The third objection refers to a customary shortcoming of Q-matrices. They are based on few individuals and generalizations about the entire population are drawn from them. In this study, the matrix is of course of more than adequate size. Furthermore our conclusions are not intended to extend to species not included in our study.

It is true that the species recorded are an eclectic sample from those extant in the world today. On the other hand we are of course dealing with a sample obtained by natural selection from the multitude of species or specieslike entities that have existed since the origin of the four genera of this study. Hypotheses regarding these extinct species will be valid only insofar as recent species reflect the course of evolutionary history.

Another point in connection with the third objection is the number of characters employed. True relationships will become apparent only insofar as the characters adequately represent the sources of variation within the species.

A fourth objection relates to the lack of equivalence in recording and interpreting the factors from the Q- and R-matrices. It compares the relative permanence of psychological tests with the relative impermanence of persons. In this study we are confronted with characters and species varying in their relative permanence, but both equally permanent when based on the time scale of the scientist investigating them.

The fifth criticism, labelling the Q-technique as descriptive rather than predictive, again is invalid when applied to the present data. Since the purpose of the study is historically descriptive and one of our aims is to divide the population of species into categories, the technique's fault for psychological research becomes a virtue in our field of investigation.

There are two evolutionary situations under which it is important to examine the two types of matrices. The first might be referred to as breakage of correlation. It occurs when in a certain evolutionary line two characters that were correlated in ancestral lines and are still correlated in related lines become independent of each



other. Under such conditions the R-matrix is a poor representation of the true relation between the two characters. There is no good way of representing such a correlation, close in one line, absent in the other. On the other hand a Q-matrix is not affected by such data.

Convergence of species for a number of characters is a second disturbing phenomenon. Here the R-matrix is not affected while the Q-matrix is affected if the convergent characters outweigh the nonconvergent ones in numbers.

We do not believe this is likely if an adequate number of characters is studied. In case of a preponderance of convergent characters and in the absence of paleontological data it is doubtful whether the systematists would be able to distinguish convergence from relationship by descent.

From a consideration of the above arguments it follows that given the objectives and material of the present study the Q-technique is to be preferred to the R-technique and the objections made by Cattell to the former method do not apply to our case. However, besides the theoretical reason for adopting the Q-technique as reflecting relationships between species there were several practical reasons for so doing. The problem of finding a suitable type of correlation coefficient between characters would have been formidable in view of the coding system adopted. Since some of the characters were present in two states only while others were present in as many as eight states, there would probably not have been any one type of correlation coefficient for all possible character combinations. A matrix based on correlation coefficients of different types would be far from desirable. Furthermore, uniformity of computational procedure was essential to efficient handling of the data by International Business Machines (IBM) equipment.

Not to be underestimated is the saving in computation resulting from adoption of a 97 x 97 species correlation matrix vs. a 122 x 122 character correlation matrix. The former requires the computation of only 4656 correlation coefficients while the latter would necessitate 7381 such coefficients.

#### *The choice of a correlation coefficient*

As a next step a suitable correlation coefficient had to be chosen to represent the correlations between species. There were serious considerations against the use of the product-moment correlation coefficient since the variables (species) are anything but normally distributed. Table 1 presents frequency distributions of state codes



TABLE 1

Frequency distributions of state codes for the characters of species 19, 56, 83 and 84.

State code	Sp. 19 f	Sp. 56 f	Sp. 83 f	Sp. 84 f
1	54	56	48	46
2	31	40	42	41
3	31	14	23	26
4	3	11	7	6
5	2	1	2	2
6				1
7	1			
$\Sigma f$	122	122	122	122

for four representative species. The distributions are highly asymmetrical. Those for species 19 and 56 approach Poisson distributions for their means when the class codes are reduced by one. Any interpretation of this agreement is dubious, however, in view of the variable number of states possible per character.

Other correlation coefficients were considered and rejected. The correlation ratio,  $\eta$ , is unsuitable since  ${}_x\eta_y$  does not necessarily equal  ${}_y\eta_x$ . Tetrachoric  $r$  would have lost some of the information available because it would necessitate reducing all characters to two states. Furthermore the theoretical assumptions of underlying normality essential to correct application of the tetrachoric correlation coefficient cannot be defended for all characters.

Another method of demonstrating an association between species would be the very simple one of counting the numbers of matches in states for the 122 characters of any pair of species of bees and then dividing this number by 122, the highest possible number of such matches. The results for species 19, 56, 83, and 84 are shown on table 2 where these "matching coefficients" are compared with product-moment correlation coefficients. The "matching coefficients" are somewhat higher than the correlation coefficients but resemble them in relative magnitude. In spite of this fact, "match-

TABLE 2

"Matching coefficients" (below diagonal) and product-moment correlation coefficients (above diagonal) between species 19, 56, 83 and 84.

	19	56	83	84
19	X	.40	.37	.37
56	.52	X	.47	.38
83	.53	.61	X	.93
84	.50	.54	.87	X



ing coefficients" were not used since they have an unknown sampling distribution, they distort resemblances by counting a 3 to 4 mismatch the equal of a 1 to 7 mismatch, and finally they would have been harder to handle by the IBM equipment available to us.

Lacking a more suitable means of correlation we adopted the product-moment  $r$ , in spite of nonnormal distribution of variates and possible heteroscedasticity. Various ways of improving the distributions by means of transformations were tried. Table 3 shows the same correlation coefficients as the upper half of the matrix of table 2, but based on  $\sqrt{X}$  and  $\sqrt{X + .5}$  transformations. The slight differences obtained do not justify the extra computational labor involved.

We have already briefly touched on the desirability of coding the data in such a way as to put all character states on the same scale. In a character with two states the code 2 indicates a situation differing greatly from one given by code 2 in a character with 7 states. This problem is also encountered in Q-matrices in psychology where the scores for different tests are often not in comparable units. This situation is usually met by normalizing the rows (tests, or in our case characters) of the raw score matrix. The authors did not per-

TABLE 3

Product-moment correlation coefficients between species 19, 56, 83 and 84 based on variates coded as  $\sqrt{X}$  (below diagonal) and as  $\sqrt{X + .5}$  (above diagonal). Compare with uncoded product-moment correlation coefficients in table 2.

	19	56	83	84
19	X	.42	.36	.37
56	.42	X	.50	.41
83	.36	.51	X	.93
84	.37	.41	.93	X

form this transformation since (1) it would have removed the common systematic factor from the matrix of correlations and would thus have lowered the correlation coefficients considerably; (2) application of the character state codes does standardize the data to a certain extent because 76 percent of the characters have either three or four states and only 3 percent have six or more states; (3) although the additional labor of normalizing the variates would not have been excessive the amount of IBM work involved in computing correlation coefficients would have been prohibitive, since a one-digit code would not have sufficed for normalized data.

The authors are well aware that their methodology of coding and correlation could profit by refinement. It is, however, our point of



view that in a pilot study of this nature such refinements are premature. Should the general method prove of value, significant results will surely emerge in spite of minor imperfections in technique.

### *Computation*

The computation of a large matrix of correlation coefficients such as the 97 x 97 bee matrix presents serious technical difficulties. Only high speed electronic computing machines are able to perform this operation with real dispatch. At the time our bee data were being processed we had only punched-card tabulating machines at our disposal. It might be noted here that a computational operation of this magnitude cannot reasonably be undertaken without some automatic computing facilities. The equipment used by the authors is that available in the University of Kansas IBM laboratory: a card punch (type 26), a verifier (type 56), an accounting machine (type 402) and a reproducing machine (type 514).

The computational problem was simplified somewhat by the fact that the variates consisted of single digits only. This increased the number of variables that the machine could process simultaneously. Each IBM card represented a character with the state code of each species for the particular character listed in separate columns. Since there are only 80 columns per card, it was impossible to record all species on any one card. A different approach was therefore adopted and the card divided as follows:

Column 1—Project code

Columns 2-4—Character code number

Column 5—Deck code (explained below)

Columns 6-8—Left blank for possible subsequent use

Columns 9-44—Multiplier columns for 36 species

Columns 45-80—Multiplicand columns for 36 species.

The 97 species were divided into group I for species 1-36, group II for species 37-72 and group III for species 73-97. Since group III used only 25 columns another 5 columns were taken up by a repetition of data on species 1 through 5, which we used as a check on computational procedure. Six decks of 122 cards each, one card per character, were then prepared. The decks were constituted as follows:

Deck	Multiplier	Multiplicand
1	Group I	Group I
2	Group II	Group II
3	Group III	Group III
4	Group I	Group II
5	Group II	Group III
6	Group I	Group III



Different card colors besides a punched code were used to distinguish the decks.

By running these decks in succession through the tabulator we were able to reduce rewiring of the board to one half of what it would have been with the minimum number of decks (3).

The method of arriving at the  $\Sigma x^2$  and  $\Sigma xy$  was the customary one of progressive digitizing with interspersed "X-cards." Running time on the 402 tabulator was some 24 hours. Punching and verifying of the cards had taken a similar amount of time. Thus the preparation of the  $\Sigma x^2$  and  $\Sigma xy$  for the entire matrix took about a week. These values were computed for a half-matrix only. However, a test deck and five test variables detected wiring errors and machine malfunction with a reasonable limit of safety.

The next step was the computation of the correlation coefficients. This was done by computers using desk calculators.<sup>4</sup> The matrix of squares and products was subdivided into manageable sections, 30 variables (species) square. All computations were checked by a different computer and, where possible, by different steps. The computational procedure employed was the customary L method.<sup>5</sup> It does not seem necessary to elaborate on the details of this method. Any good textbook of statistics will contain a section on the computation of a product-moment correlation coefficient. Furthermore, each computation center has its own setup for correlation coefficients depending on the capabilities of the machines and thus no general account need be presented here.

The correlation coefficients were calculated to four significant decimal places and entered on a matrix. Three decimal places would have been quite sufficient for this study; however four were computed in case later statistical work required greater refinement. Total computation time for this phase of the work was 160 man-hours. It should be emphasized that the time estimates given above refer to the relatively simple equipment available to us. Digital computers are now available which would handle the entire computation, from raw data to completed correlation matrix without human intervention in less than an hour. This would be only one-two hundredths of the time it took us to compute the same informa-

4. The writers at this point wish to express their appreciation to Misses Betty Becker, Marion Clyma, Jacqueline Johnson, Normandie Morrison, and Messrs. D. A. Crossley, Jr., Ralph Jones and Roger Price for their conscientious assistance with IBM work and desk computation.

5.  $r_{xy} = L_{xy} / \sqrt{L_x} \sqrt{L_y}$ , where  $L_{xy} = N\Sigma XY - \Sigma X\Sigma Y$  and  $L_x = N\Sigma X^2 - (\Sigma X)^2$ ,  $L_y = N\Sigma Y^2 - (\Sigma Y)^2$ .



tion! With every passing year electronic computers are becoming more efficient and more widely distributed. Thus the computational aspects of our method will become a progressively less important impediment.

Since the matrix of correlation coefficients was unwieldy (it also had to be subdivided into sections) and since further work with the correlation coefficients was contemplated, the latter were punched on 4656 IBM cards, one to a card. These cards were duplicated by means of the reproducing punch in order to obtain cards for a complete matrix of 9312 correlation coefficients. Information on these cards included matrix row and column numbers for the particular correlation, the coefficient with sign, and a class code for the coefficient. These class code numbers (1-22) represented 22 classes of a frequency distribution of the correlation coefficients arrayed in ascending order of magnitude with class intervals of .05. In addition, the cards contained codes for the relationship between the two species involved as evaluated by conventional systematic methods (by C. D. M.).

The correlation coefficients on punched cards have so far been put to the following uses: We have compiled a printed tape record of the full matrix, column by column, which has been very useful for reference and further computation. Another tape has been compiled giving a listing and frequency distribution of the correlation coefficients grouped in the 22 size classes. This tape has been of great value in various approaches to a classification of the relationships demonstrated by the matrix. A third tape lists the sums of the correlation coefficients, column by column. This has been necessary for the B-coefficient method briefly described below. A fourth tape presents a two-way frequency distribution showing the relation between correlation coefficients and the relationship code developed by conventional systematic methods. These tapes were prepared in a few hours running time from the correlation coefficient cards, which we still expect to use in a variety of ways.

#### *The matrix of correlation coefficients*

In the bee study the 4656 correlation coefficients computed in the above manner ranged in magnitude from  $-.0626$  for the correlation between species 26 and 92, to  $.9747$  for the correlation between species 43 and 44.<sup>6</sup> As was mentioned previously, a fre-

---

6. For lack of space the matrix cannot be reproduced here. Microfilm or IBM-tape or card copies can be obtained through the Secretary, Department of Entomology, University of Kansas, Lawrence.



quency distribution of these coefficients, grouped into 22 classes with class intervals of .05 was set up. The modal class showed a class mark of .38; this represents the most frequent class of correlation coefficients found between species in this study. However, a second mode was located at .78. This bimodality would indicate that we are dealing with two populations of correlation coefficients: those indicating close, possibly intrageneric relations and others representing more distant relations. Codes representing Michener's previous views on the relationships among the species were correlated with the above coefficients. The single correlation coefficient between the correlation matrix and Michener's codes was .80. It was encouraging to find that magnitude of the correlation coefficients in our matrix was apparently an estimate of systematic relationship as indicated by the previous classification.

Another way of examining these correlation coefficients is to study frequency distributions of the coefficients for any single species against all other species. By this means we were able to distinguish members of closely related groups of species from isolated species within a genus and these in turn from very isolated species representing monotypic genera or subgenera. For a detailed discussion and illustrations of this procedure, the reader is referred to Michener and Sokal (1957).

The absence of significant negative correlations from our matrix requires some discussion. Q-technique matrices of correlations between people (based on psychological tests) are quite likely to yield such correlations. If there are distinct, antithetical types of persons represented in the matrix, such as extroverts and introverts, it is likely that a high score for one type will be a low score for the other and vice versa. In our case evolutionary progress may be represented by either an increase or a decrease in state codes. In the majority of characters the supposedly primitive situation is an intermediate state code with two diverging evolutionary trends represented by the lower and higher code numbers. Furthermore, characters representing correlated trends were not necessarily coded along the same scale or in the same direction. It is clear that under such circumstances distantly related forms are likely to be uncorrelated rather than negatively correlated.

#### *The search for group structure*

The matrix of correlation coefficients between species can be put to a variety of uses and the analysis reported below represents merely an initial effort at an exploitation of the data. The correla-



tion coefficients serve as an absolute measure of relationship between any two species in our study, limited only insofar as the characters chosen do not represent the total correlated variation of the two species.

The search for structure among the correlation coefficients of the matrix is of course no different in aim from the search by the systematist for a natural system in an array of species. Such a system consists of a hierarchy of groups. Various methods can be used for discovering a hierarchy in data such as ours. A customary, rather simple device of the psychometrician is so-called "cluster analysis," developed to a fine art by Tryon (1939).

A concise description of the procedure (the ramifying linkage method) is given in Cattell (1944) and Thomson (1951). Because of the simplicity of the procedure, cluster methods are used extensively, although Cattell (1944, 1952) and others have pointed out that cluster analysis cannot be considered a substitute for the more involved factor analytic methods. Attempts to employ cluster analysis for finding structure in our matrix were only partially successful, since the resulting clusters were partly overlapping, *i. e.*, a given species might be a simultaneous member of two clusters. This makes good sense for intermediate forms in an abstract scheme of relationships. In a systematic hierarchic classification, however, groups at the same level have to be mutually exclusive for practical as well as for theoretical reasons, except for low level groups exhibiting reticulate evolutionary pattern (rare above the species level in animals). A further reason for the unsuitability of cluster analysis is the complexity of the clusters as more species are added to them. Although clusters are therefore not convenient in an initial search for structure, the diagram of relationships established by methods to be described below could be easily recognized in the clusters outlined by cluster analysis. A method essentially similar to cluster analysis is the  $\rho$ -group and  $\rho F$ -group method of Olson and Miller (1951) applied to three paleontological R-technique matrices. It suffers from the same drawbacks as cluster analysis.<sup>7</sup>

---

7. After the present research and manuscript had been completed one of us (R. R. S.) became acquainted with the psychometric work of Professor Louis L. McQuitty of the Michigan State University, who in recent years has developed a whole battery of refined cluster methods (McQuitty, 1955, and a series of papers in press in *The British Journal of Statistical Psychology*, *Educational and Psychological Measurement*, and *Psychological Monographs*). Several of these papers deal with psychological problems which are closely related to those of biological classification. One of the methods invented by McQuitty bears a close resemblance to our variable group method developed below. It is interesting (as well as reassuring to us) that workers in different fields had unknown to each other developed some of the same formulations. We hope to try some of McQuitty's other methods on our material. They have the advantage of simplicity and can be programmed for electronic computation without much difficulty. Indeed the time may not be far off when computation for a study such as our bee work will be a minor matter routinely handled by a computing center in a very few hours and the remaining problem will be the collection of data for the machine and the interpretation of the voluminous answers that are produced.



As a technique for grouping the species we experimented extensively with the coefficient of belonging (B-coefficient) of Holzinger and Harman (1941). It is the sum of the correlations among the members of a group divided by the sum of the correlations of these group members with the other variables (species) of the study.

Results of our B-coefficient analyses for the bees were reasonably good, as judged by the previous classification and by our subsequent investigations. There was one main drawback, however. Large species groups showed a lack of structure and relatively low B-coefficients which would make the species in these groups appear a good deal less related to one another than members of groups of two or three species. The cause of this phenomenon is not hard to find. In large species groups the denominator of the B-coefficient would include high correlations due to correlations of group members with numerous other prospective members not yet included in the group. This would tend to depress the B-coefficient values. By the time all such members have been admitted to the group, it has become so large that even the admission of a relatively unrelated variable will effect the B-coefficient only slightly.

In view of the disadvantages of the B-coefficient we developed our procedure which is presented below in a general manner together with some of the reasons for its adoption. This presentation is followed by a detailed step-by-step account of the computational procedure for readers who wish to become more familiar with it.

A nucleus of a group was established, using the two species having the highest coefficient of correlation. Then species would be added to this nucleus, one at a time, always adding first the species having the highest average correlation with members of the group. The limit of the groups could be found by decreases ( $\bar{L}_{n+1} - \bar{L}_n$ ) in the level of the average correlation  $\bar{L}_n$ , where the subscript refers to the number of members in the group. As in the B-coefficient a significant drop is empirically determined since sampling distributions of average correlations, such as  $\bar{L}_n$ , are unknown. By developing first lower groups (species groups), then by the same method grouping these into larger groups (sometimes subgenera), and these into still larger or higher groups, etc., it has been possible to develop a hierarchy of groups for which the diagram of relationships (figure 1) can serve as a representative. Each number in this figure represents a different species; for a list of the species concerned see Michener and Sokal (1957).

Since  $\bar{L}_n$  is not amenable to rigorous statistical treatment it was decided to recompute correlation coefficients (using Spearman's sum



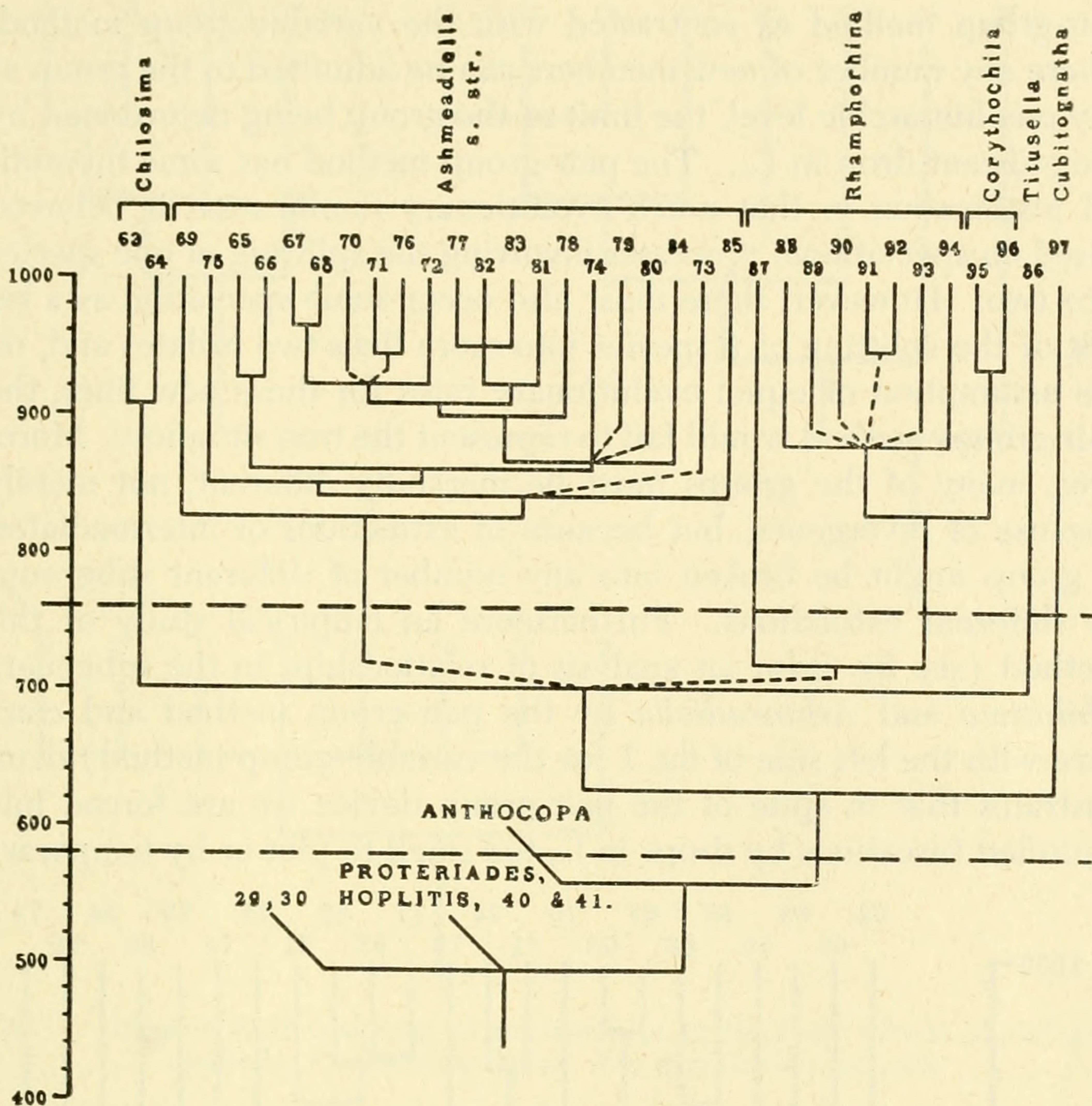


FIG. 1. Diagram of relationships for the genus *Ashmeadiella* obtained by the weighted variable group method. Ordinate: magnitude of correlation coefficient multiplied by 1000. Exact correlations between any two joining stems can be found by reading the value on the ordinate corresponding to the horizontal line connecting the stems. This value becomes approximate and maximal in cases of multifid furcations. Broken lines used where more than three stems join are for convenience only; the horizontal connecting line has the same significance as elsewhere. "Roofs" over species numbers at the summits of the lines delimit subgenera containing more than one species, as based on C. D. M.'s previous findings and not on this study. Generic names are in small capitals. The horizontal broken lines are not relevant to the present account; they are explained by Michener and Sokal (1957).

of variables method) after the group limits at each hierarchic level had been reached. Thus we returned at the end of each grouping procedure to a new matrix of correlation coefficients about which confidence statements might be made. Two further considerations in the final choice of a method for grouping remain to be mentioned:

We might have admitted only one new member for each group at a given hierarchic level, thus obtaining a diagram of relationships consisting of bifurcations only. We have called this method the



pair-group method as contrasted with the variable-group method, where any number of new members can be admitted to the group at any one hierarchic level, the limit of the group being determined by a significant drop in  $\bar{L}_n$ . The pair-group method has some theoretical justification in that much evolutionary ramification is believed based on speciation processes involving the splitting of one species into two. However, there must also occur some speciation as a result of the splitting of a species into more than two isolates and, on the assumption of equal evolutionary rates for these new lines, the pair-groups method would fail to represent the true situation. Moreover, many of the groups must be markedly different, not merely because of divergence, but because of extinctions of intermediates. A group might be broken into any number of different subgroups by different extinctions. Furthermore an empirical study of this method (see fig. 2 for an analysis of relationships in the subgenera *Chilosima* and *Ashmeadiella* by the pair-group method and compare with the left side of fig. 1 for the variable-group method) demonstrates that in spite of the pair-group device we are forced into multified furcations by drops in  $\bar{L}_n$  too small to plot or by temporary

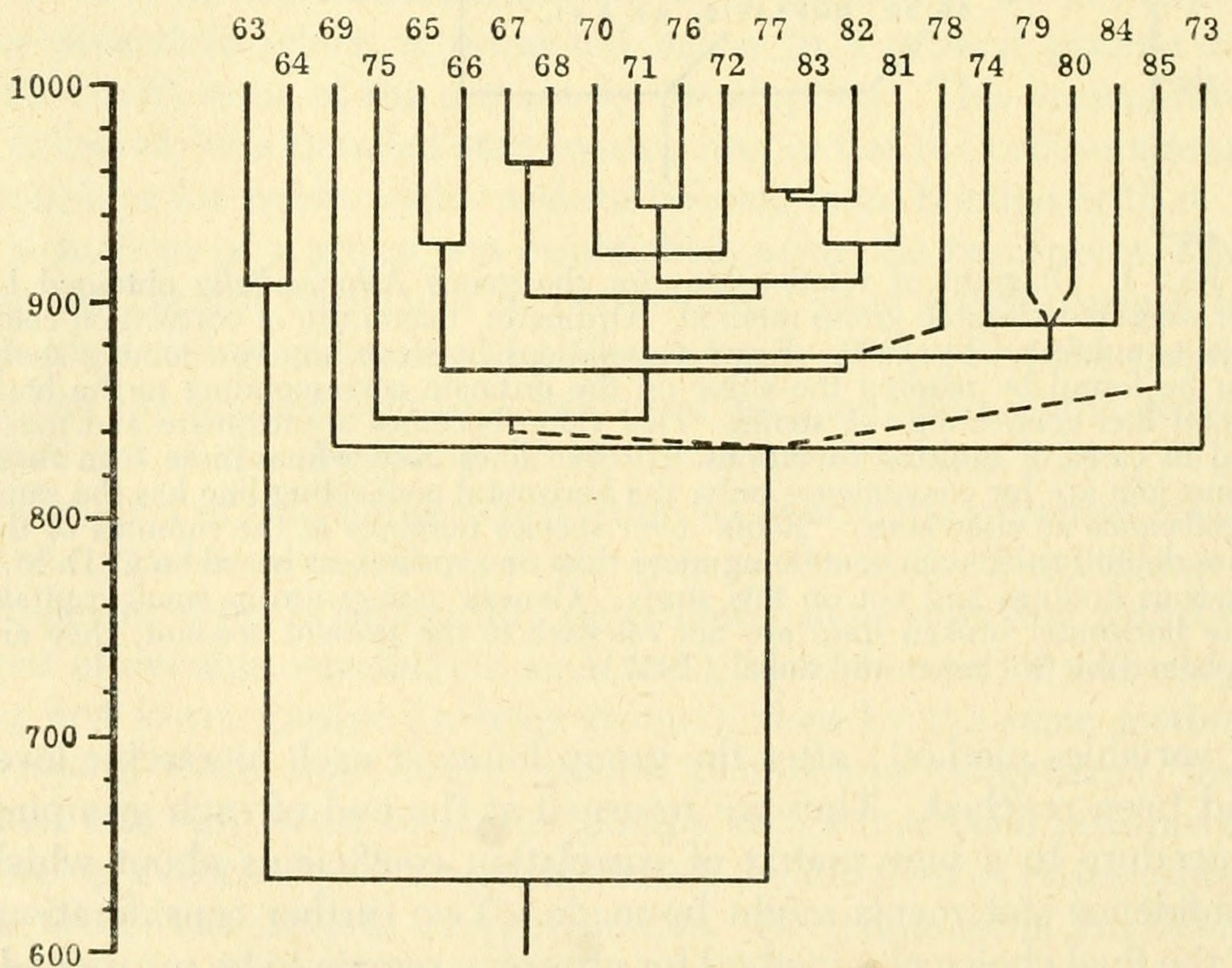


FIG. 2. Diagram of relationships for the subgenera *Chilosima* (63-64) and *Ashmeadiella s. str.* (65-85), obtained by the method of pair-groups, *i. e.* diagrams would ideally consist of bifurcations only. Stems have been weighted. Explanatory comments as for figure 1.



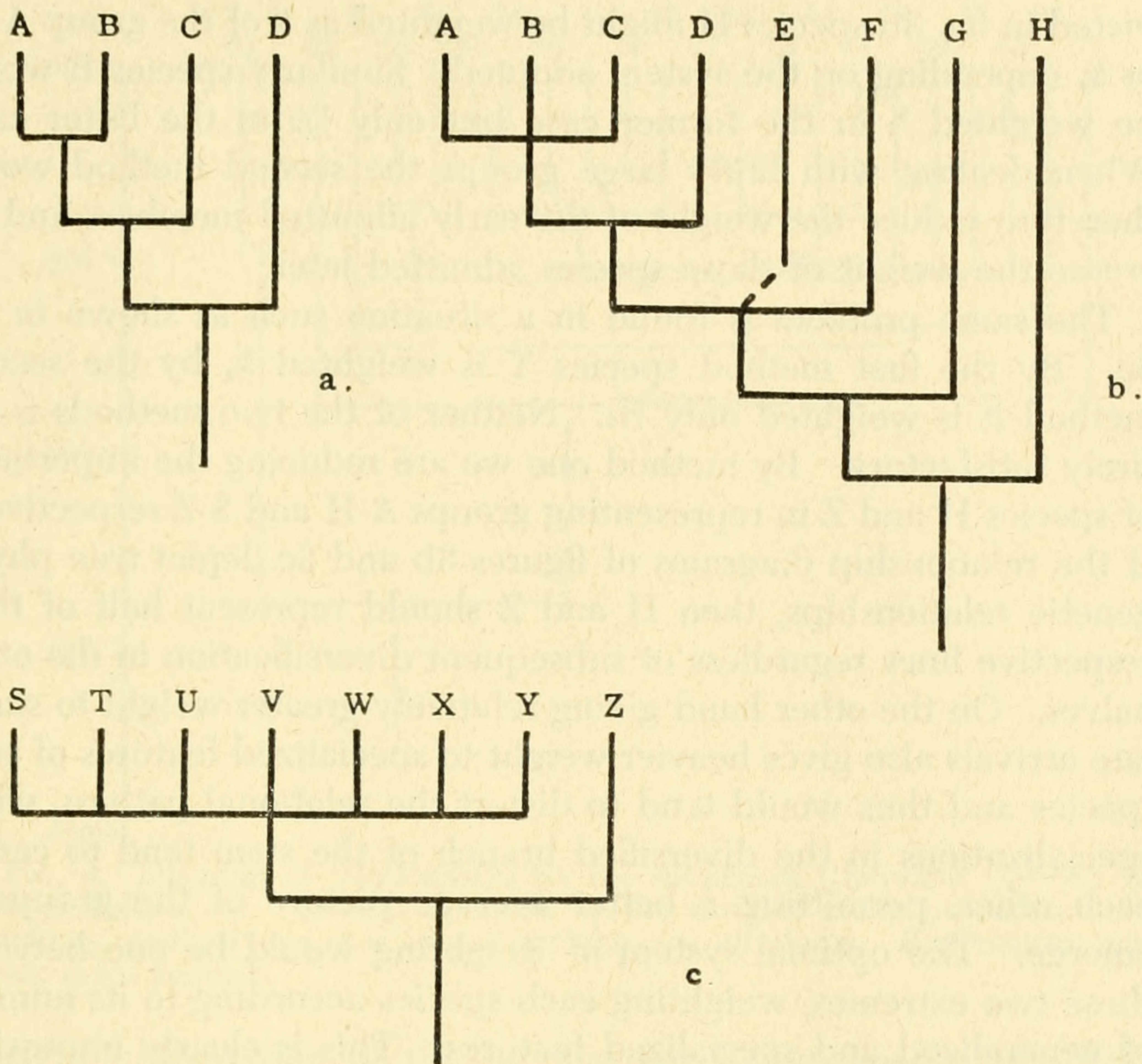


FIG. 3. Hypothetical diagrams of relationships to illustrate effects of different methods of weighting stems. For explanation see text.

reversals of  $\bar{L}_n$  values, discussed below. Thus the variable-group method was adopted as the more reasonable and flexible of the two.

A second consideration is how to weight the variables during the recalculation of the correlation matrix after each grouping procedure. A simple diagram (fig. 3a) will make this issue clear. A and B represent the two species with the highest correlation coefficient. The  $\bar{L}_n$  for C against A and B is significantly below  $r_{ab}$ , so that A and B are represented as being closer to each other than they are to C. When studying the relation of a fourth species D with group ABC we face the following problem: Should we calculate the correlation of ABC against D with A, B and C equally weighted or should we weight  $A = B$  and  $AB = C$ ? Rephrased biologically, the problem is whether to relate species D with the homogeneous group ABC, or with the stem AB-C, where C carries as much weight in determining the relation with D as do A and B together. Although in a simple case, such as the one described above the two alternatives may not produce very different results, in a situation such as de-



picted in fig. 3b species H might be weighted as  $\frac{1}{8}$  of the group A-H, or  $\frac{1}{2}$ , depending on the system adopted. Similarly species B would be weighted  $\frac{1}{8}$  in the former case but only  $\frac{1}{2}$  in the latter case. When dealing with fairly large groups the second method would therefore reduce the weight of the early admitted members and increase the weight of those species admitted later.

The same problem is found in a situation such as shown in fig. 3c. By the first method species T is weighted  $\frac{1}{8}$ , by the second method it is weighted only  $\frac{1}{14}$ . Neither of the two methods is entirely satisfactory. By method one we are reducing the importance of species H and Z in representing groups A-H and S-Z respectively. If the relationship diagrams of figures 3b and 3c depict true phylogenetic relationships, then H and Z should represent half of their respective lines regardless of subsequent diversification in the other halves. On the other hand giving relatively greater weight to single late arrivals also gives heavier weight to specialized features of such species and thus would tend to distort the relational pattern, while specializations in the diversified branch of the stem tend to cancel each other, permitting a better average picture of the groups to emerge. The optimal system of weighting would be one between these two extremes, weighting each species according to its number of generalized and specialized features. This is clearly impossible without renewed introduction of a subjective element into our procedure. We therefore adopted the second method, *i. e.*, the weighting of new members as equal to the sum total of all old group members, thinking it to be the less objectionable of the two. We feel that this method will represent stems more correctly and that bias introduced by specializations of late joiners will be kept down by the large number of characters considered in our study.

We are reassured in our decision by the results of a comparative study on the subgenera *Chilosima* and *Ashmeadiella*. Figure 4 shows the results of a variable-group analysis of these subgenera by weighting method one, while results by method two can be seen in the left side of the diagram of fig. 1. General agreement as to relationships and level of furcations is very good. The main difference between the two diagrams is that in method one group 77-81<sup>8</sup> first receives 79 before receiving group 67-72, 78 and 84, while in method two it first receives group 67-72, then 78 and then 79 among others.

---

8. In the interest of brevity groups will be identified by their leftmost (in the diagram) and rightmost members with a dash separating the two. Thus 77-81 means group 77, 82, 83, 81. It clearly does *not* include all species ranging in number from 77 through 81, and includes some beyond that numerical range.



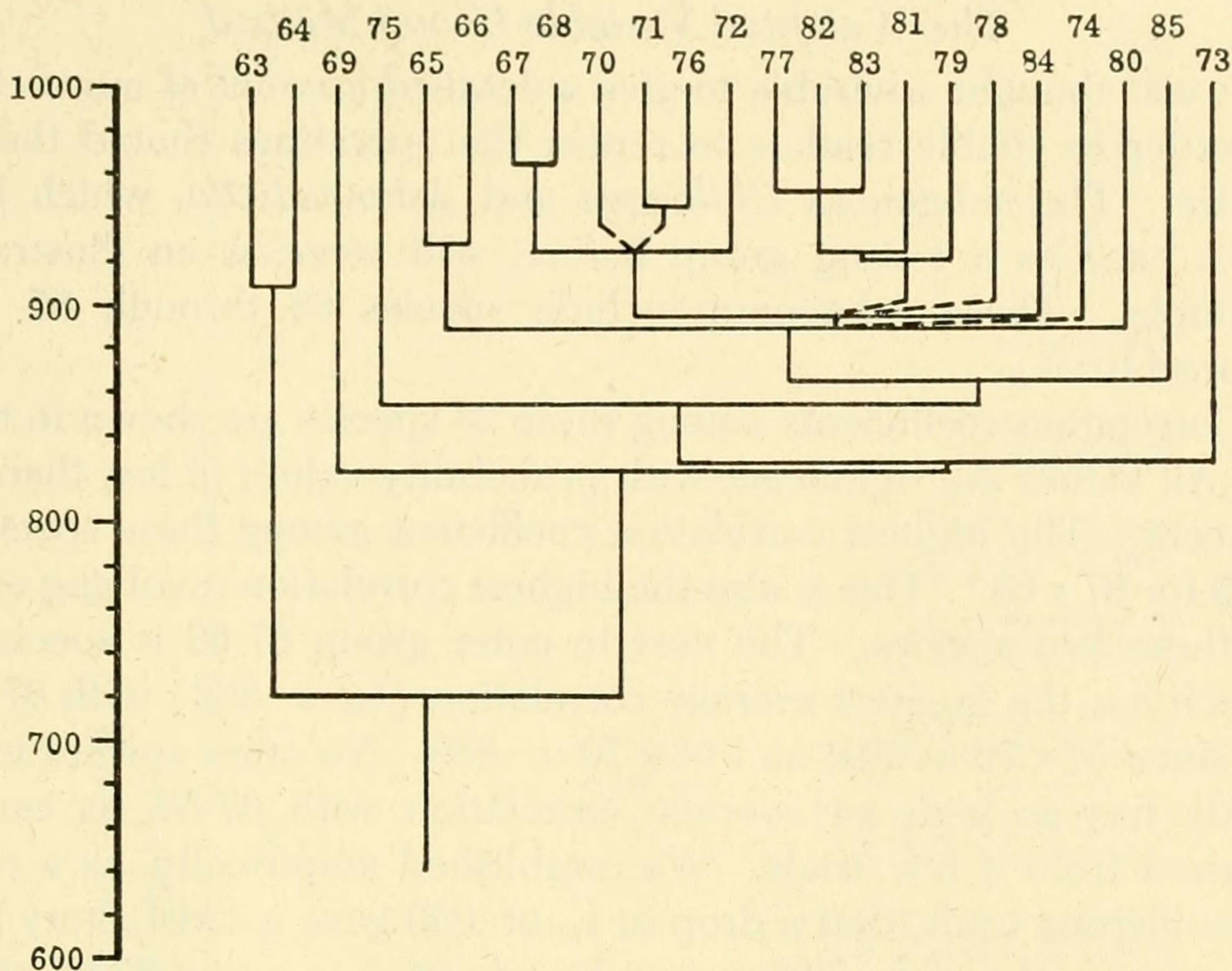


FIG. 4. Diagram of relationships for the subgenera *Chilosima* (63-64) and *Ashmeadiella s. str.* (65-85), obtained by the variable group procedure under weighting method one, *i. e.*, equal weights for all stems. Explanatory comments as for figure 1.

Careful examination of the original correlation coefficients makes the reasons for these differences clear. Group 77-81 is closer to 79 than to 78 except that 81 is closer to 78 than to 79. Also 81 is closer to 67-72 than are 77, 82, and 83. Therefore in method two, where 81 receives as much weight as 77, 82 and 83 together, 67-72 joins the nuclear group first. This is also partly due to the fact that unequal weighting of the species in the 67-72 group favors those close to the 77-81 group. Since 78 is closer to 67-72 than 79, the latter, while originally quite close to 77-81 is now temporarily delayed and 78 joins the combined group 67-81 before 79 does. These relations are at too low a phyletic level to be included in the original diagram of relationships drawn by C. D. M. who feels that there is little that can be obtained from classical systematic studies of these species to suggest whether method one or method two is preferable. In view of the small over-all differences between the two methods and especially in view of the fact that the lines concerned all join by either method with a difference in correlation coefficients of less than .06, it may well be that we have made too great an issue of the matter. In any case we feel confident that weighting method two will present us with a reasonably bias-free picture.



*The Weighted Variable Group Method*

It was thought advisable to give a detailed account of our method in order to enable readers to repeat the operations should they so desire. The subgenera *Chilosima* and *Ashmeadiella*, which have been used as a testing group before, will serve as an illustrative example. These subgenera include species 63 through 85 (see figure 1).

Correlation coefficients among these 23 species are shown in table 4. All values are significant with probability values of less than one percent. The highest correlation coefficient among these species is .965 for  $\overline{67 \times 68}$ .<sup>9</sup> This is also the highest correlation involving either of these two species. The next to enter group 67-68 is species 70 which has the greatest average correlation ( $\bar{L}_n = .892$ ) with 67 and 68 since  $\overline{67 \times 70} = .896$  and  $\overline{68 \times 70} = .889$ . No other species in the study has as high an average correlation with 67-68, as can be learned from a few trials. We established empirically, as a result of numerous trials, that a drop in  $\bar{L}_n$  of .030 gave a satisfactory limit for groups; therefore 70 is not to be admitted to group 67-68 at this particular time. Another high correlation involving species other than 67 and 68 is  $\overline{77 \times 83} = .951$ . This is also the highest correlation for the two species concerned. Next to join this nucleus is species 82 with an  $\bar{L}_n$  value of .936. The drop is less than .030; therefore 82 is admitted. Next to join is species 79 with an  $\bar{L}_n$  value against 77, 82 and 83 of .905. There is now a significant drop from the previous  $\bar{L}_n$  value and 79 is excluded for the time being. Drops in  $\bar{L}_n$  are always measured from the previous  $\bar{L}_n$ , not from the initial  $\bar{L}_n$ . Our second group is therefore 77-83. In a similar manner we established groups 63-64, 65-66 and 71-76, each consisting of only two species.

So far only 11 species out of the 23 of the study have been placed into groups. A systematic survey was then made of the remaining 12 species to see if any group had been missed. For example, examination of species 69 revealed that its highest correlation was with species 72 ( $\overline{69 \times 72} = .820$ ). However, this latter value was not the highest correlation for 72, since  $\overline{72 \times 76} = .904$ . Thus 72 might eventually join the group containing 76, and 69 might join the group containing 72, both of which events came to pass at a later stage of the analysis. At the present time, however, species 69 and 72 are left unattached to any group. Similarly the remaining ten species in the study were shown not to belong to any nuclear group. To

9. We shall use this symbolism in place of the more formal  $r_{67-68}$



TABLE 4—Matrix of Correlation Coefficients Among Species 63 to 85  
(Extracted from larger matrix of 97 species)

	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	
63	XXX																						
64	.908	XXX																					
65	.652	.653	XXX																				
66	.624	.623	.927	XXX																			
67	.692	.681	.806	.771	XXX																		
68	.682	.672	.812	.767	.965	XXX																	
69	.523	.511	.725	.676	.775	.817	XXX																
70	.752	.718	.860	.832	.896	.889	.772	XXX															
71	.685	.725	.831	.793	.864	.858	.728	.897	XXX														
72	.669	.710	.859	.843	.849	.870	.820	.889	.902	XXX													
73	.560	.611	.699	.620	.739	.719	.706	.727	.779	.789	XXX												
74	.582	.637	.803	.731	.770	.758	.697	.748	.805	.779	.823	XXX											
75	.569	.611	.786	.777	.737	.754	.771	.712	.771	.847	.683	.796	XXX										
76	.697	.741	.837	.809	.883	.868	.734	.922	.945	.904	.751	.791	.752	XXX									
77	.566	.608	.826	.784	.804	.833	.764	.780	.809	.861	.767	.862	.838	.798	XXX								
78	.677	.719	.762	.718	.814	.792	.704	.848	.893	.811	.772	.762	.693	.895	.802	XXX							
79	.567	.610	.830	.763	.748	.763	.768	.728	.764	.817	.772	.832	.826	.769	.926	.800	XXX						
80	.643	.605	.822	.769	.787	.757	.716	.783	.828	.786	.789	.865	.802	.800	.863	.822	.852	XXX					
81	.635	.679	.822	.788	.827	.847	.776	.842	.873	.866	.805	.834	.780	.872	.885	.883	.857	.816	XXX				
82	.617	.661	.853	.783	.793	.816	.787	.821	.859	.888	.809	.857	.827	.850	.925	.844	.897	.855	.923	XXX			
83	.589	.631	.866	.821	.815	.819	.738	.823	.861	.855	.762	.887	.821	.843	.951	.838	.891	.894	.907	.948	XXX		
84	.542	.584	.794	.747	.725	.738	.659	.705	.740	.741	.667	.845	.766	.737	.886	.725	.852	.849	.788	.851	.926	XXX	
85	.555	.605	.759	.695	.790	.819	.800	.774	.787	.854	.832	.843	.763	.784	.829	.723	.778	.762	.859	.849	.827	.744	



set up one of the latter we required a correlation coefficient which was the highest one for both participating species (*i. e.*, the reciprocally highest correlation).

After the groups had been delimited a new correlation matrix was computed considering the newly formed groups as single variables, *i. e.*, the previous matrix of 23 variables (matrix 1) was reduced to one of 17 variables (matrix 2). It is self-evident that the only correlation coefficients in need of recomputation were those involving new groups. Correlations involving only species that had remained single were not altered in any way. As a matter of fact, a procedure was devised by means of which the correlations were not even recopied, but variables joined into groups were crossed out and the new group variables were entered along the margins of the old matrix. The actual computational procedure is quite simple and considerably less complicated than the computations for finding the original correlation coefficients. It is described in the paper by its originator (Spearman, 1913) and also by Holzinger and Harman (1941). Let us illustrate this method by computing the correlation between groups  $(63)_1$  and  $(67)_1$ <sup>10</sup>. The general formula for this computation is

$$r_{q.Q} = \frac{\square qQ}{\sqrt{q + 2\Delta q} \sqrt{Q + 2\Delta Q}}$$

where  $\square qQ$  is the sum of all correlations between members of one group with the other group,  $\Delta q$  is the sum of all correlations between members of the first group,  $\Delta Q$  is a similar sum between members of the second group,  $q$  is the number of species in group one and  $Q$  the number of species in group 2. Thus in this particular case  $\square qQ$  equals  $(63 \times 67) + (63 \times 68) + (64 \times 67) + (64 \times 68) = .692 + .682 + .681 + .672 = 2.727$ ;  $\Delta q$  in this case equals only  $(63 \times 64) = .908$  while  $\Delta Q$  equals  $67 \times 68 = .965$ , since each of these groups consists of two species only. In cases where a group consists of 3 species, for example, the  $\Delta$  term consists of the sum of  $(1 \times 2) + (1 \times 3) + (2 \times 3)$ . In the present case  $q = Q = 2$  species. Substituting into the formula given above:

$$\frac{(63)_1 \times (67)_1}{\sqrt{2 + 2(.908)} \sqrt{2 + 2(.965)}} = \frac{2.727}{\sqrt{2 + 2(.908)} \sqrt{2 + 2(.965)}} = .704$$

10. The notation  $(63)_1$ , refers to the group of species formed in matrix 1, the lowest numbered member of which is species 63, *i. e.*, to group 63-64. Similarly  $(67)_1$ , refers to 67-68, and  $(77)_1$ , to 77-82-83.



These computations can be set up in a systematic manner and are then neither particularly complicated nor time consuming. In the special case where we wish to calculate the correlation coefficient between a single species ( $x$ ) and a new group ( $q$ ), the formula is amended as follows:

$$r_{x,q} = \frac{\Sigma r_{x,q}}{\sqrt{q + 2\Delta q}}$$

An illustration is the correlation of species 69 with group  $(77)_1$ .  $\Sigma r_{x,q}$  equals  $(69 \times 77) + (69 \times 82) + (69 \times 83) = .764 + .788 + .738 = 2.290$ , while  $\Delta q = (77 \times 82) + (77 \times 83) + (82 \times 83) = .925 + .951 + .948 = 2.824$ .

$$\text{Then } \underline{69 \times (77)_1} = \frac{2.290}{\sqrt{3 + 2(2.824)}} = .779$$

In such a manner a new 17 x 17 correlation matrix (matrix 2) was constituted. From this point on the species groups  $[(63)_1, (65)_1, (67)_1, (71)_1, \text{ and } (77)_1]$  were tested as though they were single species.

Once matrix 2 had been computed the identical grouping procedure was followed. Group  $(71)_1$  had a mutually highest correlation with species 70 at .923. They were then joined by 72 and group  $(67)_1$  at  $\bar{L}_n$  levels of .903 and .885 respectively. These affiliations of 72 and  $(67)_1$  were also their highest correlations. The next prospective joiner was species 81 at  $\bar{L}_n = .859$ , *i. e.* not quite the established drop of .030. However species 81 had highest relations not with the previous species but with group  $(77)_1$  with a correlation of .923. Therefore it was excluded from consideration as a candidate for the earlier group and the runner up, species 78, used instead. The latter gave an  $\bar{L}_n$  value of .844, clearly a significant drop from .885. Species 81 meanwhile was used in a nucleus of a new group 77-81  $[(77)_2]$ . Situations such as the above were the exception. In general the relations and choices were entirely straightforward and could be left to the discretion of the computing assistants.

At the end of each grouping procedure the remaining single variables were checked to avoid missing groups with low correlations between members. With each grouping procedure the matrix of correlation coefficients became smaller and the job of recomputation less. The weighting procedure adopted by us was automatic in that all correlation coefficients used were from the previous matrix and not the initial one. It took eleven matrices to obtain a single



group out of the 23 species of the two subgenera. This amount of work could have been reduced by raising the minimum recognized difference in  $\bar{L}_n$  level above .030 but there would have been a resulting loss of detail in the diagram of relationships. Conversely, however, reducing the recognized difference below .030 would not have increased the meaningful detail, since even .030 was too small to prevent the occasional reversal of  $r$  values discussed below.

Once computed, the relations were represented as diagrams of relationships as in figure 1. The ordinate at the left of each diagram is graduated in units of  $1000 \times r$ . The correlations between any joining stems in the diagram can be read by measuring the level along the ordinate of the horizontal line connecting the stems. Thus species 63 and 64 are correlated at a level of .908, while group 63-64 is related to group 67-72 at .702. Furcations involving more than three lines are shown by broken lines converging on the midpoint of the horizontal line as in group 88-96 of the above figure. The tops of the figures are at a level of 1000 (correlation of 1) since obviously each species is perfectly correlated with itself.

In cases of groups of only two stems the  $\bar{L}_n$  level corresponds to the correlation coefficient of the two stems. When more than two stems join to form a group the highest  $\bar{L}_n$  level was graphed for all group members. Thus while  $77 \times 83$  equals .951,  $\bar{L}_n$  for 82 against 77 and 83 equals .936. The group of these three species (77-83) is shown related at levels .951. Occasionally the correlation coefficients for the same group in successive matrices will rise a little. Thus in this same figure groups 63-64, 69-85, 87-96, and species 86 are shown joining at .702. The first three groups actually joined at .671, but species 86 which joined their group at the next matrix did so at level .702. This type of situation, which occurred infrequently, might lead one to express concern about the validity of the method, since regular decreases in levels of correlation coefficients and  $\bar{L}_n$  values are expected. However, it can be shown from Spearman's formula for the correlation of sums of variables that slight increases in the levels of correlation coefficients of the sums of variables above the correlations of their component variables are possible. For example, if A and B have formed the nucleus of a group at  $r_{a,b} = .9$  and C is about to join them, then by the rules of the variable group method both  $r_{a,c}$  and  $r_{b,c}$  must  $< r_{a,b} = .9$ . It can then be shown that  $r_{(ab),c}$  must be  $< .925$ . Thus  $r_{ab,c}$ , while it will usually be  $< r_{a,b}$ , could be slightly more than .9. Similar situations can be shown to exist with larger-sized groups. The increases found by us were well



below the mathematically possible limits. In all such cases the relations were represented as multifid furcations of all the stems involved in the reversal and at the highest of the several  $\bar{L}_n$  levels considered.

In a successful method of studying relationships, the results of the analysis should be relatively independent of the number of species in the correlation matrix. If at least one species per species-group is included in a matrix, the ideal method of analysis should reproduce the diagram of relationships based on an earlier study of a larger matrix. If the method can be shown to produce similar results, the fact that our matrix contains only a sample from the population of species can be ignored with greater assurance.

We tested this question by subjecting the odd-numbered species in the entire genus *Ashmeadiella* to a weighted variable group analysis. This should give an adequate cross-section of relationships in that genus. Since some trends might be lost by exclusion of the

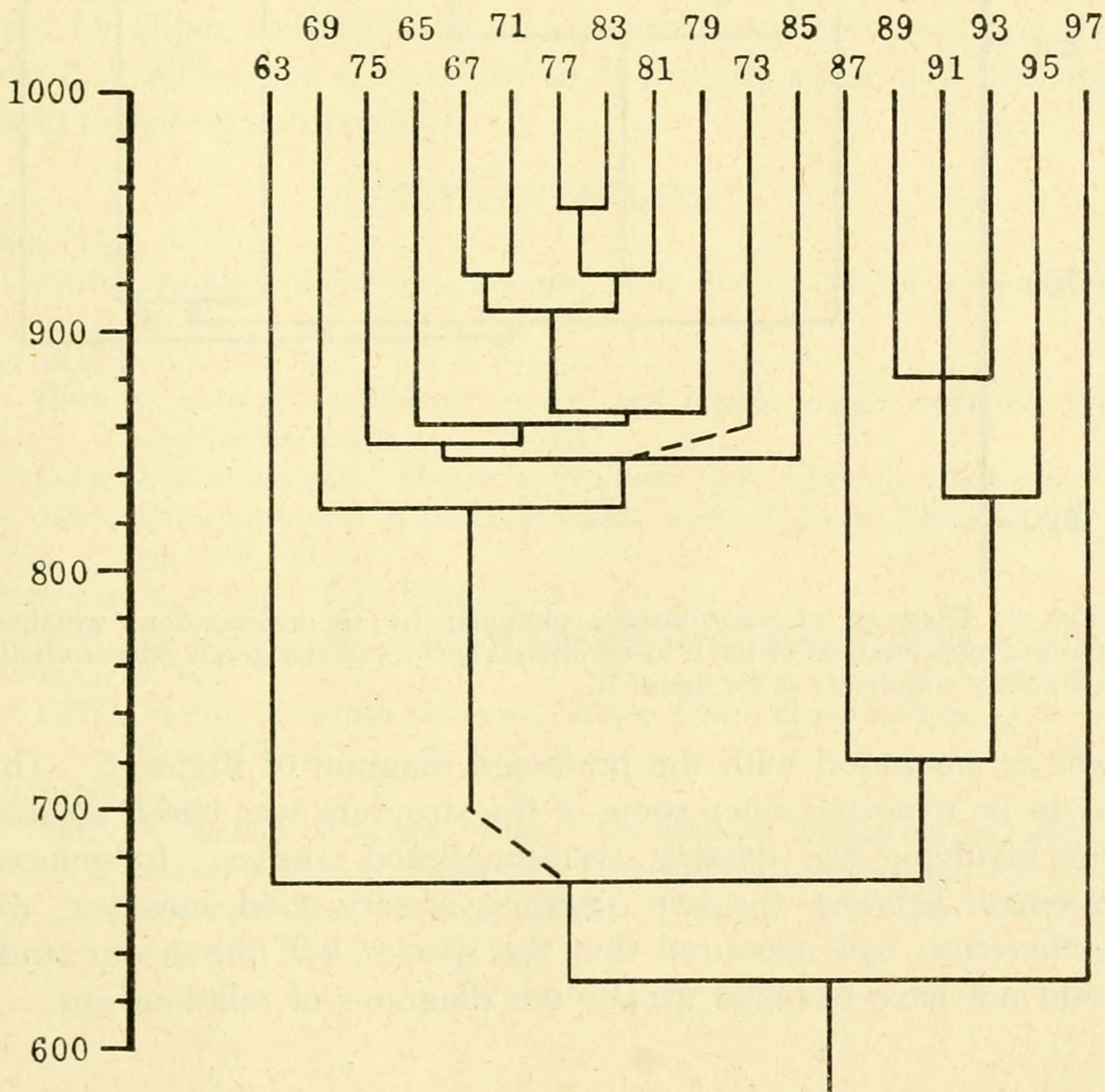


FIG. 5. Diagram of relationships predicted for odd numbered species of the genus *Ashmeadiella* on the basis of the relationship diagram of figure 1. Explanatory comments as for figure 1.



even-numbered species a special diagram of relationships was prepared from figure 1 by using only odd-numbered species. Figure 5 shows this predicted diagram of relationships.

Figure 6 shows the results of the weighted variable group analysis on the odd-numbered species. There is less structure in this dia-

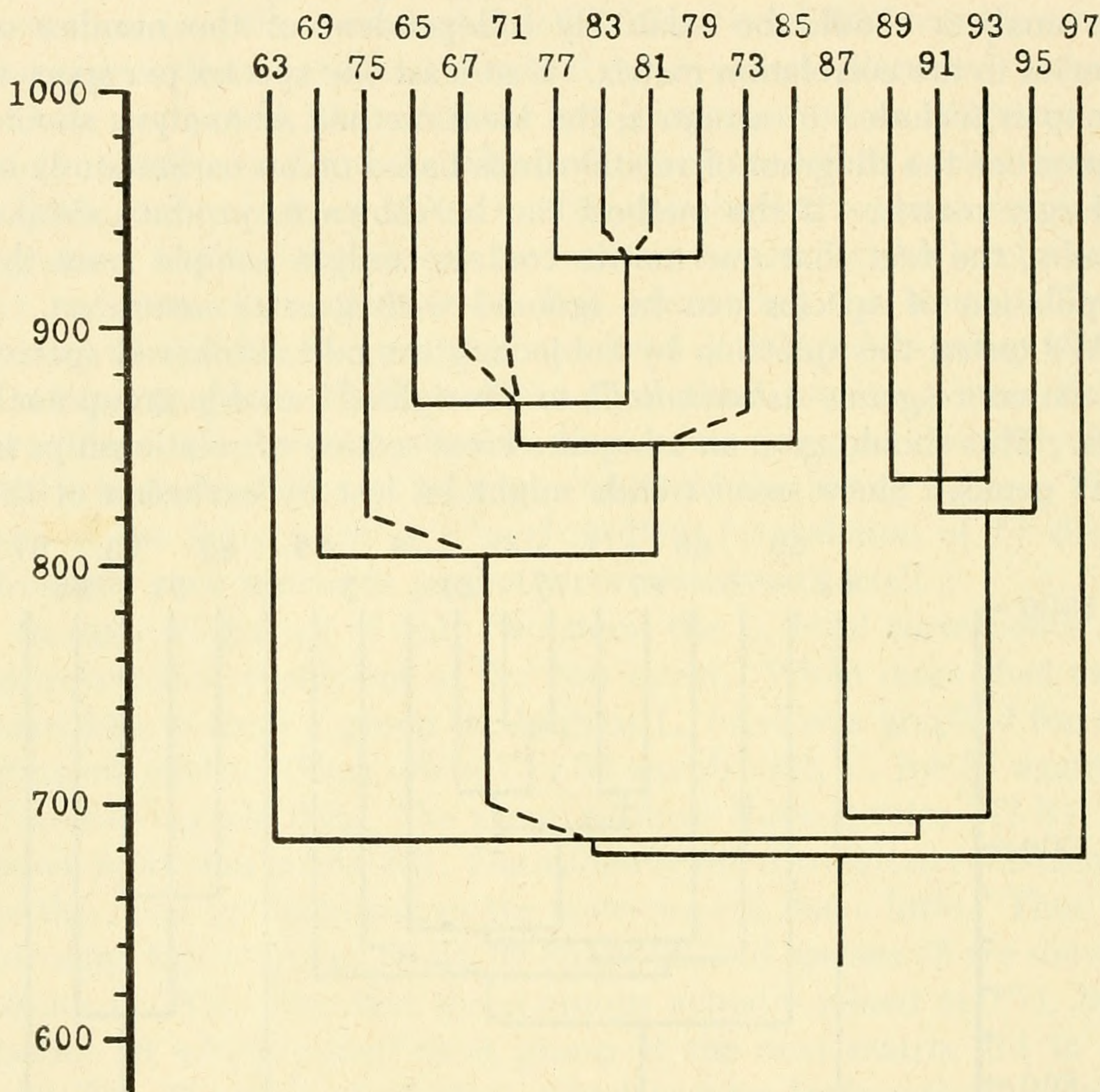


FIG. 6. Diagram of relationships obtained by an independent weighted variable group analysis of the odd-numbered species of the genus *Ashmeadiella*. Explanatory comments as for figure 1.

gram as compared with the predicted diagram of Figure 5. This was to be expected since some of the structure was based on relations involving the missing even-numbered species. In general, agreement between the two diagrams is very good, however, and we therefore feel reassured that the species left out in our study would not have changed greatly our diagrams of relationships.



## CONCLUDING COMMENTS

A detailed discussion of the comparisons of our findings with the previous classifications of the four genera of bees is given by Michener and Sokal (1957). It will suffice here to state that general agreement was good but that a number of taxonomic re-evaluations seemed necessary as an outcome of these analyses. It should be remembered that while diagrams such as figure 1 may suggest phylogenies, in reality they only indicate static relationships. As indicated in the paper referred to, additional refinements were devised to give diagrams of relationships which we believe more nearly approach phylogenetic trees.

In view of these results we are encouraged to believe that, since the methods we have described are increasingly practical with the growing availability of high speed computers, this or similar schemes will be more widely utilized with different groups of organisms. Although the method we have described is a first attempt and would profit by either simplification or refinement, we believe it is a step toward reducing the subjectivity of systematic work, and therefore a step in the right direction.

## LITERATURE CITED

- BURT, C. L.  
1937. Correlations between persons. *Brit. Journ. Psychol.*, vol. 28, pp. 59-96.
- CATTELL, R. B.  
1944. A note on correlation clusters and cluster search methods. *Psychometrika*, vol. 9, pp. 169-184.  
1952. *Factor analysis*. Harper & Brothers, New York, pp. xiii + 462.  
1954. Growing points in factor analysis. *Austral. Journ. Psychol.*, vol. 6, pp. 105-140.
- HOLZINGER, K. U., and H. H. HARMAN.  
1941. *Factor analysis*. Univ. of Chicago Press, Chicago, pp xii + 417.
- HOWELLS, W. W.  
1951. Factors of human physique. *Amer. Journ. Phys. Anthrop.*, n. s., vol. 9, pp. 159-191.
- MCQUITTY, L. L.  
1955. A method of pattern analysis for isolating topological and dimensional constructs, Research Report AFPTRC-TN-55-62, Air Force Personnel and Training Center, Lackland Air Force Base, San Antonio, Texas, v + 38 pp.



MICHENER, C. D., and R. R. SOKAL.

1957. A quantitative approach to a problem in classification. *Evolution*, vol. 11, pp. 130-162.

OLSON, E. C., and R. L. MILLER.

1951. A mathematical model applied to a study of the evolution of species. *Evolution*, vol. 5, pp. 325-338.

SOKAL, R. R.

1952. Variation in a local population of *Pemphigus*. *Evolution*, vol. 6, pp. 296-315.

1958. Quantification of systematic relationships and of phylogenetic trends. Proc. Tenth Internat. Congress Entomology, Montreal, (in press).

SPEARMAN, C.

1913. Correlations of sums or differences. *British Journ. Psychology*, vol. 26, pp. 344-361.

STEPHENSON, W.

1936. The inverted factor technique. *Brit. Journ. Psychol.*, vol. 26, pp. 344-361.

STROUD, C. P.

1953. An application of factor analysis to the systematics of *Kaloterme*s. *Systematic Zool.*, vol. 2, pp. 76-92.

STURTEVANT, A. H.

1942. The classification of the genus *Drosophila* with descriptions of nine new species. *Univ. Texas Publ.*, no. 4213, pp. 1-51.

THOMSON, G. H.

1951. The factorial analysis of human ability. 5th ed. Houghton Mifflin Company, New York, pp. xv + 383.

TRYON, R. C.

1939. Cluster analysis. Edwards Bros., Ann Arbor, Mich., pp. 1-122.