Hunt Institute for Botanical Documentation
5th Floor, Hunt Library
Carnegie Mellon University
4909 Frew Street
Pittsburgh, PA 15213-3890
Telephone: 412-268-2434
Email: huntinst@andrew.cmu.edu
Web site: www.huntbotanical.org

The Hunt Institute is committed to making its collections accessible for research. We are pleased to offer this digitized item.

Usage guidelines
We have provided this low-resolution, digitized version for research purposes. To inquire about publishing any images from this item, please contact the Institute.

Statement on harmful and offensive content
The Hunt Institute Archives contains hundreds of thousands of pages of historical content, writing and images, created by thousands of individuals connected to the botanical sciences. Due to the wide range of time and social context in which these materials were created, some of the collections contain material that reflect outdated, biased, offensive and possibly violent views, opinions and actions. The Hunt Institute for Botanical Documentation does not endorse the views expressed in these materials, which are inconsistent with our dedication to creating an inclusive, accessible and anti-discriminatory research environment. Archival records are historical documents, and the Hunt Institute keeps such records unaltered to maintain their integrity and to foster accountability for the actions and views of the collections' creators.

Many of the historical collections in the Hunt Institute Archives contain personal correspondence, notes, recollections and opinions, which may contain language, ideas or stereotypes that are offensive or harmful to others. These collections are maintained as records of the individuals involved and do not reflect the views or values of the Hunt Institute for Botanical Documentation or those of Carnegie Mellon University.

About the Institute
The Hunt Institute for Botanical Documentation, a research division of Carnegie Mellon University, specializes in the history of botany and all aspects of plant science and serves the international scientific community through research and documentation. To this end, the Institute acquires and maintains authoritative collections of books, plant images, manuscripts, portraits and data files, and provides publications and other modes of information service. The Institute meets the reference needs of botanists, biologists, historians, conservationists, librarians, bibliographers and the public at large, especially those concerned with any aspect of the North American flora.

Hunt Institute was dedicated in 1961 as the Rachel McMasters Miller Hunt Botanical Library, an international center for bibliographical research and service in the interests of botany and horticulture, as well as a center for the study of all aspects of the history of the plant sciences. By 1971 the Library's activities had so diversified that the name was changed to Hunt Institute for Botanical Documentation. Growth in collections and research projects led to the establishment of four programmatic departments: Archives, Art, Bibliography and the Library.

18 April 1968

Frank G. Hawksworth
Rocky Mountain Experiment Station
240 W. Prospect Street
Fort Collins, Colo. 80521

Dear Frank:

I received our manuscript and have looked it over. Your suggestion for a title page is good. I will fill in our half of the footnote. The minor changes you suggest are fine with me. I have a few additional suggestions myself. Whenever you are ready to get together for the final polishing, give me a ring.

Very truly yours,

George F. Estabrook

GFE:gm

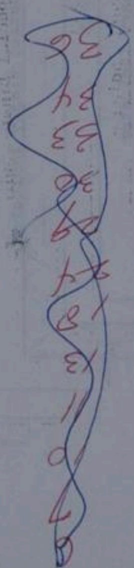
Parameter card

32

6

26

10
0
16
7



Hawksworth - Arcuella

Ch		K			
1	8	1	17	2	0
2	3	1	18	3	0
3	8	1	19	7	1
4	9	1	20	6	1
5	6	1	21	5	1
6	3	0	22	6	1
7	2	0	23	6	1
8	9	1	24	6	1
9	6	1	25	4	1
10	3	0	26	4	1
11	4	1	27	2	0
12	4	1	28	10	1
13	4	1	29	15	0
14	4	1	30	3	0
15	4	1	31	5	1
16	4	1	32	7	1

Hawksworth - Arceuthobium 2

obs #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
1	3	2	1	6	1	2	1	1	1	1	1	1	1	2	2	1	1	2	0	2	0	6	3	4	3	3	1	1	F	3	5	2		
2	3	1	1	5	1	2	1	1	1	1	1	1	1	1	2	1	1	2	0	2	0	6	2	4	3	3	1	1	F	3	5	3		
3	4	1	2	5	1	2	1	1	1	1	1	1	2	2	2	2	1	2	3	1	0	6	5	6	4	3	1	1	F	3	5	3		
4	4	1	1	6	2	2	1	1	1	1	1	2	2	1	2	1	1	2	0	2	0	6	2	4	3	3	1	1	F	3	5	2		
5	4	0	1	7	2	2	1	1	1	1	1	1	2	1	2	2	1	2	0	2	0	6	2	4	3	3	1	1	F	3	5	0		
6	3	1	1	5	1	2	1	1	1	1	0	1	0	0	0	0	1	2	3	0	0	6	0	0	0	0	1	1	F	3	5	4		
7	4	3	2	6	2	2	1	1	1	1	0	1	0	0	0	0	1	2	3	0	0	6	4	4	4	4	1	1	F	3	5	4		
8	4	3	3	7	2	3	1	4	3	3	2	3	0	0	0	0	2	1	4	0	0	6	4	4	4	4	2	1	D	3	5	7		
9	4	2	3	7	2	3	1	0	4	3	2	3	2	2	2	2	2	1	4	0	0	6	4	3	4	4	2	1	D	3	5	6		
10	4	3	2	5	2	3	1	4	4	3	2	3	0	0	0	0	2	1	3	0	0	6	4	0	4	4	2	1	D	3	5	0		
11	4	4	2	6	3	3	1	4	3	3	2	3	0	0	0	0	2	1	3	0	0	6	4	0	4	4	2	1	D	3	5	0		
12	4	1	2	5	2	3	1	7	2	3	4	3	2	2	2	2	2	1	3	0	0	6	4	0	4	4	2	1	D	3	5	4		
13	4	2	2	6	2	3	1	6	3	3	2	3	0	0	0	0	2	1	3	0	0	6	4	0	4	4	2	1	D	3	5	0		
14	4	2	2	6	2	3	1	7	3	3	2	3	2	1	2	2	2	1	4	0	0	6	4	0	4	4	2	1	D	3	5	0		
15	2	3	1	3	1	3	1	2	2	2	2	2	1	2	2	2	2	1	0	0	0	6	6	1	2	2	2	1	E	3	3	0		
16	2	2	1	2	1	3	1	2	2	2	2	2	1	2	2	2	2	1	0	0	0	6	6	1	2	2	2	1	E	3	2	0		
17	2	0	1	2	1	3	1	2	2	2	1	2	2	1	2	2	2	1	0	0	0	6	6	2	3	3	2	1	E	3	2	0		
18	2	3	1	2	1	3	1	2	2	2	0	1	0	0	0	0	2	1	0	0	2	6	6	2	3	3	2	1	E	3	2	0		
19	2	0	1	2	1	3	1	2	2	2	0	1	0	0	0	0	2	1	0	0	0	6	6	3	3	3	2	1	E	3	2	0		
20	2	3	1	2	1	3	1	2	2	2	1	1	1	1	2	2	2	1	0	2	0	6	6	3	3	3	2	1	E	3	2	0		
21	2	3	1	2	1	3	1	2	2	2	1	1	2	1	2	2	2	1	0	0	0	6	6	2	3	3	2	1	E	3	2	0		
22	2	0	1	2	1	3	1	2	2	2	0	1	0	0	0	0	2	1	0	0	0	6	6	2	4	4	2	1	E	3	2	0		
23	2	0	1	2	1	3	1	2	2	2	0	1	0	0	0	0	2	1	0	0	0	6	6	2	4	4	2	1	E	3	2	0		
24	2	0	1	2	1	3	1	2	2	2	0	1	0	0	0	0	2	1	0	0	0	6	6	2	4	4	2	1	E	3	2	0		
25	3	3	2	6	2	3	1	4	2	3	2	1	1	1	2	1	2	1	3	0	5	4	3	2	4	2	2	1	2	3	4	4		
26	5	0	2	8	2	3	1	7	2	3	2	2	2	1	2	2	2	1	3	3	4	4	2	2	4	2	2	1	2	3	5	5		
27	4	3	2	5	2	3	1	4	2	3	2	2	2	2	2	2	2	1	3	3	4	4	2	2	4	2	2	1	2	3	5	5		
28	6	3	2	8	2	3	1	5	2	3	2	2	2	1	2	2	2	1	3	3	4	4	2	2	4	2	2	1	2	3	5	5		
29	3	3	2	7	2	3	1	5	2	3	1	0	0	0	0	0	2	1	3	3	4	4	2	2	4	2	2	1	2	3	5	5		
30	4	3	3	8	2	3	1	3	2	3	0	2	2	2	2	2	2	1	3	3	4	4	2	2	4	2	2	1	2	3	5	5		
31	4	3	2	7	2	3	1	4	2	3	1	2	1	2	2	2	2	1	3	3	4	4	2	2	4	2	2	1	2	3	5	5		
32	4	3	2	7	2	3	1	5	2	3	2	2	1	0	2	2	2	1	3	3	4	4	2	2	4	2	2	1	2	3	5	5		
33	4	0	2	9	2	3	1	4	2	3	0	1	0	2	2	2	2	1	3	3	4	4	2	2	4	2	2	1	2	3	5	5		
34	3	2	1	6	2	2	1	1	1	1	1	1	2	2	2	2	1	2	0	2	0	6	3	4	3	3	1	1	F	3	5	2		
35	4	0	2	6	2	2	1	1	1	1	1	2	2	2	2	2	1	2	0	2	0	6	3	4	3	3	1	1	F	3	5	2		
36	4	0	2	7	2	2	1	1	1	1	1	2	4	1	3	2	1	2	0	2	0	6	4	4	3	3	1	1	F	3	5	0		
37	4	2	3	7	3	3	1	5	4	3	2	3	3	2	3	2	2	1	3	3	4	4	2	2	4	3	2	1	2	3	5	4		
38	4	3	3	7	3	3	1	7	4	3	2	3	3	2	3	3	2	1	3	3	4	4	2	2	4	3	2	1	2	3	5	4		
39	5	0	3	8	3	3	1	7	4	3	2	2	2	2	3	2	2	1	3	3	4	4	2	2	4	3	2	1	2	3	5	5		
40	4	3	3	8	3	3	1	5	4	3	2	3	3	2	3	2	2	1	3	3	4	4	2	2	4	3	2	1	2	3	5	5		
41	4	3	3	6	2	3	1	4	3	3	2	3	3	2	3	2	2	1	3	3	4	4	2	2	4	3	2	1	2	3	4	4		
42	4	2	2	5	2	3	1	7	3	3	2	3	3	2	3	2	2	1	3	3	4	4	2	2	4	3	2	1	2	3	4	4		
43	4	2	3	5	2	3	1	5	2	3	2	3	3	2	3	2	2	1	3	3	4	4	2	2	4	3	2	1	2	3	4	4		
44	4	2	3	8	2	3	1	4	2	3	2	3	3	2	3	2	2	1	3	3	4	4	2	2	4	3	2	1	2	3	4	4		
45	4	2	3	8	2	3	1	0	0	3	0	0	0	0	0	0	2	1	3	3	4	4	2	2	4	3	2	1	2	3	4	4		
46	4	3	3	6	2	3	1	0	0	3	0	0	0	0	0	0	2	1	3	3	4	4	2	2	4	3	2	1	2	3	4	4		
47	4	1	3	9	3	3	1	4	0	3	0	0	0	0	0	0	2	1	3	3	4	4	2	2	4	3	2	1	2	3	4	4		
48	4	1	3	6	2	3	1	0	0	3	0	0	0	0	0	0	2	1	3	3	4	4	2	2	4	3	2	1	2	3	4	4		
49																																		
50																																		

APPLICATION OF AN INFORMATION THEORY MODEL FOR CHARACTER ANALYSIS IN THE GENUS ARCEUTHOBIUM (VISCACEAE).[1]

Frank G. Hawksworth[2], George F. Estabrook[3] and David J. Rogers[3]

1 Paper No. 20 from the Taximetrics Laboratory, University of Colorado, Boulder, Colorado. This research was supported by (1) a cooperative aid agreement between the U. S. Forest Service, Rocky Mountain Forest and Range Experiment Station, Fort Collins, Colorado, and the University of Colorado and (2) N.I.H. grant GM 13974-01.

2 U. S. Forest Service, Rocky Mountain Forest and Range Experiment Station, Colorado State University, Fort Collins, Colo. 80521.

3 Taximetrics Laboratory, Department of Biology, University of Colorado, Boulder, Colorado 80302.

OUTLINE

ABSTRACT

A character for a group under study embodies the notion of "similar with respect to a basis for comparison" by inducing a partition of the group into classes called character states. Objects which are members of the same state are considered similar. The information in a character can be measured. This enables us to determine that

i. Some characters share information with other characters.

ii. Some properties may be predicted from knowledge of other properties.

iii. Conventional notions of character "equality" and "weighting" are often misleading.

iv. In a very real sense, a classification is a character and may be measured for its information preserving capacity.

The information in selected characters of the genus Arceuthobium was analysed in this manner; the results of this analysis are presented.

I. Introduction

Arceuthobium is a highly specialized but clearly defined genus of mistletoe family (Viscaceae). These dwarf mistletoes are small, leafless parasites of conifers. They cause serious damage, both in growth reduction and mortality, in western coniferous forests. The genus occurs throughout much of the Northern Hemisphere although the main concentration of taxa is in western North America.

The genus is characterized by several distinctive features: lack of leaves, explosive fruits, no central vascular cylinders in the stems, a chlorophyllous endosperm, a ring-like archesporium, and exclusive occurrence on Pinaceae. Although the generic limits of the group are well-defined, the taxonomic status of various members of the genus has long been in doubt.

George Engelmann was the first taxonomist to study intensively the genus Arceuthobium. In papers dating from 1849 to 1880 he described several new North American species and varieties but, unfortunately, he did not complete his planned monograph of the group. Occasional new taxa were added to the genus in the early 1900's but no comprehensive treatment of the genus appeared until Gill's (1935) paper on "Arceuthobium in the United States." Gill reduced the number of species in the United States from 13 to 5 with two of them embracing several host forms. With the increasing recognition of Arceuthobium as destructive parasites, the expanded biological interest in the genus, and the difficulties encountered with the application of Gill's host-form concept, (e.g. Kuijt, 1960; Hawksworth and Graham, 1963) the need for additional systematic work on the genus became apparent. Thus, for the past several years, Dr. Delbert Wiens, of the University of Utah, and the senior author have been working toward a taxonomic revision of Arceuthobium in

North America. A preliminary paper on the genus in Mexico has been published (Hawksworth and Wiens 1965).

Much of the present confusion in the systematics of Arceuthobium is a result of the extreme morphological reduction that these plants have undergone as a result of their parasitic habit. Commonly used morphological features such as leaves, stems, and flowers, are either lacking or so reduced that they are of limited value in distinguishing members of the group. Thus, it was apparent that there was a need for studies of a wider range of characters, in addition to morphological ones.

Since 1962, Dr. Wiens and the senior author have made numerous field trips throughout North America to study the dwarf mistletoes and we have studied all taxa (except A. bicarinatum of Hispaniola) in their living state. These trips have resulted in over 1,200 collections which are filed in the Forest Pathology Herbarium of the Rocky Mountain Forest and Range Experiment Station, Fort Collins, Colorado. These specimens plus the examinations of several thousand others at the major North American Herbaria form the basis for our systematic studies.

Many types of taxonomic data have been assembled: morphology of shoots, flowers, and fruits; hosts and host reactions; pollen characteristics; karyology; phenology of flowering and fruit maturity; and chromatography of shoot pigments. These studies resulted in large amounts of data, the taxonomic significance of which could best be determined by computer techniques.

Fortunately, the character analysis program based on the information theory model of Estabrook and Rogers (Estabrook 1967) was found to be ideally suited to the analyses of these Arceuthobium data. The purpose of the character analysis program is to determine the taxonomic value of characters. It is desirable to have data on the information contained in the characters and

character states, the characters which are correlated, and the best character construction to serve the purpose of providing a natural classification, before a computer classification program is employed.

Results of the character analysis program on 34 characters of the following types are presented here.

1. Shoots. Size, color, types of branching, etc. (10 characters).
2. Flowers. Type of inflorescence, flower size, etc. (10 characters).
3. Pollen. Size, exine thickness, height of spines (5 characters).
4. Fruits. Size, glaucousness (2 characters).
5. Phenology. Time of meiosis, flowering, seed dispersal, fruit maturation period (4 characters).
6. Hosts. Principal hosts, consistency and type of witches' brooms (3 characters).

Although 34 characters were used in this study, only the following fourteen that are cited as examples of the character analysis technique are described here. These 14 were selected because they were of unusual interest (some are classical characters in the genus, some new) or because they best exemplified various features of the character analysis program. The complete list of characters will be published later in a taxonomic revision of Arceuthobium by F. G. Hawksworth and Delbert Wiens.

5. Width of Third Shoot Segment

1. Up to 1.0 mm
2. 1.1 - 2.0 mm
3. 2.1 - 3.0 mm
4. 3.1 - 4.0 mm
5. 4.1 - 5.0 mm
6. 5.1 - 6.0 mm

6. Branching Type

1. No accessory branches
2. Accessory branches sometimes verticillate
3. Accessory branches always flabellate

7. Sexual Dimorphism

1. Branching similar
2. Branching different

9. Width of Pre-flowering Lateral Staminate Spikes

1. None
2. 1 mm
3. 2 mm
4. 3 mm
5. 4 mm
6. 5 mm

12. Diameter of Staminate Flower

1. 2.0-2.4 mm
2. 2.5-2.9 mm
3. 3.0-3.4 mm
4. Over 3.5 mm

14. Location of Anther (Distance from Tip of Lobe)

1. 0.3-0.4 mm
2. 0.5-0.6 mm
3. 0.7-0.8 mm

16. Width of Staminate Lobes

1. 0.6-0.9 mm
2. 1.0-1.3 mm
3. 1.4-1.7 mm

17. Pistillate Flowers

1. Verticillate
2. Opposite

20. Peak Flowering Period

1. March-April
2. May-June
3. July
4. August
5. September
6. October or later

25. Pollen Spine Length

1. 1.0-1.4 μ
2. 1.5-1.9 μ
3. 2.0-2.5 μ

27. Shoot Color

1. Pistillate and staminate shoots of same color
2. Pistillate and staminate shoots of different color

28. Color of Pistillate Shoots

1. Green
2. Yellow
3. Orange
4. Red
5. Purple
6. Brown
7. Black

29. Principal Hosts

1. Abies (concolor, grandis)
2. Abies (magnifica)
3. Abies (other species)
4. Pseudotsuga
5. Picea (mariana, glauca)
6. Picea (engelmannii, pungens)
7. Larix
8. Tsuga
9. Pinus (Subsect. Cembroides)
10. Pinus (Subsect. Leiophyllae)
11. Pinus (Subsect. Australes)
12. Pinus (Subsect. Ponderosae)
13. Pinus (Subsects. Sabinianae and Oocarpae)
14. Pinus (Subsect. Contortae)
15. Pinus (flexilis, aristata)
16. Pinus (strobiformis)
17. Pinus (lambertiana)

34. Flowering Group (Wiens, 1968)

1. Group I (spring flowering, spring meiosis)
2. Group Ia (winter flowering, fall meiosis)
3. Group II (summer flowering, summer meiosis)
4. Group III (spring flowering, fall meiosis)

We employ a new concept in this study, the idea that a classification itself is an information-carrying system, and that in this sense, the subgeneric and supra-specific categories in the classification may be considered as the states of a character. Thus, the classification is a new character (No. 35) and its states are (see Fig. 1): Group I (1 taxon - abietis-religiosae), Group II (1 taxon - verticilliflorum), Group III (3 taxa -

americanum, douglasii, pusillum), Group IV (1 taxon - strictum), Group V (6 taxa - vaginatum complex, gillii complex, globosum), Group VI (12 taxa - campylopodum complex, rubrum). This concept allows an analysis of the information contained in the classification in relationship to the descriptive characters used to establish the classification. The purpose of this type of analysis is to give a more specific knowledge of both the characters and the information content of the classification. See page 17 for further discussion.

Fig. 1 shows a sub-generic classification established on the basis of the 34 descriptive characters, using the Graph Theory Model of clustering (Wirth, Estabrook and Rogers, 1966, and Irwin and Rogers, 1967). This as yet unpublished classification is based on a sample of the best 200 specimens representing all the known North American taxa of the genus. This is the first classification which gives a well-structured sub-generic and supra-specific arrangement of the taxa of Arceuthobium. The classification has not been published because we are awaiting inclusion of chromatographic and other data. The final classification may be slightly altered by this additional information, but we anticipate that the general picture will be unchanged. In any case, for this paper we are primarily interested in the classification as an example.

(Insert Figure 1)

II. Information in a Character

A character for a group of organisms* (such as the genus Arceuthobium, for example) provides information about the similarities and differences between pairs of objects in the group with respect to some basis for comparison independently of other bases for comparison. If a means exists for deciding when a pair of objects from the group is similar or dissimilar with respect to a designated basis for comparison, and the notion "is similar to" is transitive ("is similar to" will be transitive if knowing that A is similar to B, and B is similar to C always allows us to conclude that A is similar to C as well) then a character may be formed. The relation "is similar to" allows us to divide the group of objects described into subgroups in the following way. Choose A among the objects in the group. Place together with A all other objects to which A is similar. This will constitute A's subgroup. Now choose an object, B, that is not in A's subgroup. Form B's subgroup. Continue in this way to construct a "partition" for the original group. Now two objects are similar if they belong to the same subgroup and different if they belong to different subgroups. These subgroups are called the states of the character. It is now possible to associate with each state a summarizing

---

* Characters may also be defined for groups of groups of organisms, e.g. a group of populations, or for groups of groups of groups of organisms, in the case of Genera within a family, etc. In this more general situation the basic concept of character remains unchanged. The word "object", rather than organism, or groups of organisms, will be used throughout to represent the more general situation.

description of the basis for comparison as represented by the membership in that state. States will be identified by this description or by a numerical symbol which will stand for the description. Characters 6, 34, and 35 are good examples of the generality of this approach to characters.*

The information in a character constructed in this way may be described in two ways which we shall call Quality and Quantity.

i. Quality. This is a consideration of how reliable the notion of "similar" and "different" embodied in a character actually reflects biological fact. It is possible to define characters for a group which are characters structurally but which do not reflect biological fact. Clearly, such characters should be avoided whenever possible. In some cases this is easy. Few (if any) competent taxonomists would choose to say that two organisms were similar if the color of their respective herbarium labels were the same, and different otherwise. On the other hand, in virtually no case can we assert with confidence that biological reality has not been done some injustice by the character purporting to represent it. Thus the quality of a character is reflected in the states that it gives rise to, the membership in those states, and the reasons for that membership. The extent to which a character reflects biological reality depends on the professional acuity of the biologist who defines it. There are many considerations which he can bring to bear on his decisions. Among these are:

1. Experimental evidence suggesting such things as genetic plasticity and ontogenetic differentiation.

2. Field observations suggesting such things as environmental selection pressures and competition stresses.

3. History of the group, reflecting what other specialists have thought.

---

* It should be clear from these discussions that the traditional distinction between qualitative and quantitative characters is not applicable to the present concept of character

4. His own personal judgement (or prejudice) as a professional biologist. It is assumed that the professional will endeavor to construct characters of high quality.

ii. Quantity. We must assume that the characters defined by the professional in whose aid this technique was developed are of high quality. This may not always be the case, but as the characters have been professionally endorsed we will proceed to measure the quantity of information they possess with no further considerations of quality.

The quantity of information in a character is determined by two factors:

1. the number of states to which the character has given rise;

2. the distribution (in number of objects) of the objects over the states.

The mathematical manner in which quantity is defined is given by Estabrook (1967). A mathematical discussion is inappropriate here. The process may be described intuitively as follows:

The quantity of information in a character is a measure of how difficult it is (on the average) to guess the state membership of a randomly chosen object. The units of this measure may be thought of as the average number of "yes-no" questions required by the guesser to ascertain this state membership. This may seem, at first, a strange measure for information. However, one should think of "yes-no" questions as the acquisition of sufficient information to describe an unknown with respect to the basis of comparison in question. Perhaps the measure will seem more natural than

It may now be made clear how the above two factors influence the quantity of information in a character.

1. As the number of states in a character increases, the guesser has more choices to choose from, and hence it becomes more difficult for him to make his choice.

2. If one state is known to contain more objects than another, it is more likely that a randomly chosen unknown will belong to the one than to the other; thus it is easier to guess.

It should be clear, in the light of these two factors, that different characters contain different quantities of information and that this difference is unrelated to any quality differences that may exist. Thus we see that there is an inherent inequity among characters with respect to quantity as well as quality. These considerations should be borne in mind when any assertions that characters are "equal" or "unweighted" are made.

iii. Examples. As shown in the following tabulation, the quantity of information increases with the number of states in the character. (See Table I.)

(Insert Table I here)

The example below shows the effects of distribution on the quantity of information. (See Table 2.)

(Insert Table 2 here)

Pollen spine length (char. 25), which is more nearly equally distributed through the three states, has more than twice as much information as branching type (char. 6) which has a highly skewed distribution. Branching type has long been used as a principal taxonomic character in this group (Gill, 1935) and it is useful in separating out a few taxa. For example, two species each segregate out in State 1 (no accessory branches) and State 2 (branching verticellate) but the bulk of the taxa are in State 3 (branching flabellate). Thus, while the character is of value in segregating a few members, the above analyses indicate that, for the rest of the genus, the character is not useful in distinguishing taxa. Characters with a highly skewed distribution within states have low information quantity, but this does not imply that they are low in quality.

III. Information Shared by Two Characters

1. General Discussion. The quantity of information in a character is a measure of how difficult it is to predict the state membership for that character of an object randomly chosen from the study. It is interesting to ask how that measure is affected if we (the guessers) are provided with

additional information about some other character (namely, the state membership of the randomly chosen object for that other character). Let us suppose that in trying to guess the state membership for character 6 (branching type) for a random object in the study, we are always told to which state of character 7 (sexual dimorphism) that object belongs. If there were a dependence between characters 6 and 7, we would expect this additional information to help us guess branching type, and in general after learning this additional information about sexual dimorphism, fewer yes-no questions would be required (on the average). If such a dependence did not exist, then we would not expect this additional information to be of any help. Since we may always ignore this additional information if we choose, the measure of information in character 6 is never greater than the measure of information left in character 6 after being given the additional information about 7. Let H(6) represent the amount of information in character 6 and let H(6/7) represent the amount of information left in character 6 when we are also presented with information about character 7. Let us denote the difference H(6) - H(6/7) with the symbol R(6,7); read the redundancy of 6 and 7. If the additional information about character 7 did not help us guess about character 6, then H(6) = H(6/7) and R(6,7) = 0; thus $R(6,7) \geq 0$.

The number R(6,7) measures the amount of information which characters 6 and 7 share in common; thus $H(6) \geq R(6,7)$. The number H(6/7) measures the amount of information in character 6 that is not also in character 7. Thus, we have R(6,7) = R(7,6) and H(7) - H(7/6) = R(7/6) = R(6,7) = H(6) - H(6/7).

To provide a geometrical analog for these concepts, the "heuristic information space" was derived (Estabrook, 1967). This is a rectangle whose area corresponds to information. We may represent the information in a character with the area enclosed by a circle drawn in this rectangle.

(Insert Fig. 2)

If we draw two characters on this space, the following correspondences exist:

(Insert Fig. 3)

The information relations for pairs of characters can be described and classified with the aid of the heuristic information space:

(Insert Fig. 4)

If two characters are virtually independent, their circles overlap very little. For example, sexual dimorphism (char. 7) and shoot color (char. 27) share very little common information.

(Insert Fig. 5)

If two characters share information their circles will overlap. The two characters, width of third shoot (char. 5) and peak flowering period (char. 20), have about the same amount of information and share about one-third of this information.

(Insert Fig. 6)

An example of shared information in two characters having unequal amounts of information is found in branching type (char. 6) and pistillate flowers (char. 17). About two-thirds of the information in char. 17 is included in char. 6.

(Insert Fig. 7)

If one character is a refinement of another, the one merely further divides the states of the other. The one will contain more information than the other and the heuristic information space looks like:

(Insert Fig. 8)

ii. Prediction. In instances for which the redundant or shared information for a pair of characters constitutes a large portion of the information in one member of the pair, it is possible to predict the state membership for that member from knowledge of the other member. For this a conditional probability table is constructed. The probability that a random object belongs to one of the states of character 7 (across the top of Table 3) on the condition that it is known to belong to a given state of character 34 (down the side of Table 3) is found in the table at the intersection of the appropriate row and column. E. g., the probability of an object being in state 2 of character 7 given that it already belongs to state 3 of character 34 is 0.019.

Sexual dimorphism (char. 7, 2 states, information 0.431) and flowering group (char. 34, 4 states, information 1.695) share common information of 0.360. Thus about 83% of the information in char. 7 is included in char. 34.

(Insert Table 3 here)

no F

If an object is ~~classed~~ placed in flowering states 1, 2, or 4, it will be possible to predict its placement in the appropriate sexual dimorphism state with absolute certainty. If it is placed in flowering state 3, its probability of correct sexual dimorphism placement is almost certain.

Conversely, sexual dimorphism (char. 7) and shoot color (char. 27) share so little common information that the possibility of predicting shoot color from sexual dimorphism is almost nil.

(Insert Table 4 here)

The conditional probabilities of the two states of char. 7 are virtually the same as the unconditional probabilites Therefore, information about shoot color does not affect our prior knowledge of sexual dimorphism.

iii. Distance. It is now possible to define a measure of distance or closeness for pairs of characters. This measures how similar the information in each character is to the other. Define:

$$D(X,Y) = \frac{H(X/Y) + H(Y/X)}{H(X/Y) + H(Y/X) + R(X,Y)}$$

for any pair of characters X and Y.

D(X,Y) = 0 means that X coincides with Y.

D(X,Y) = 1 means that X and Y are independent.

Intermediate values reflect intermediate conditions. It might be noted that

this measure of distance is very reluctant to drop significantly below its maximum of one unless the two characters X,Y show definite interdependencies. Values of distance below .5 or so generally correspond to pairs of characters so interdependent that it will be apparent to the biologist even before his information is processed by a machine. It was found, as was to be expected, that with the present study many character pairs enjoyed distances close to one and the overwhelming majority of character pairs gave rise to distances greater than .7. Unlike Redundancy which is an absolute measure of the information in common for two characters, Distance is a bounded measure and must always be between 0 and 1.

iv. The Overall Relations of a Character. Measures have been used to compare all pairs of characters for the present Arceuthobium study. Overall distance is the average distance taken over all the other characters; similarly, overall redundancy. Characters that are relatively unrelated to the other characters are shown by high overall distance and low overall redundancy.

Character	Overall Distance	Overall Redundancy
Anther Location (Char. 14)	0.9694	0.0765
Staminate Lobe Width (Char. 16)	0.9560	0.1039

This indicates that these two characters are relatively unrelated to the other characters used. Anther location was first used by Engelmann (1850) to separate A. campylopodum from A. robustum but recent work shows that anther location is of little taxonomic value. Staminate lobe width has not been traditionally used and our work shows that it, too, was basically unrelated to the other characters for this study.

Characters that are highly related to most other characters are shown by relatively low overall distance and high overall redundancy.

Character	Overall Distance	Overall Redundancy
Principal Hosts (Char. 29)	0.7617	0.9044
Flowering Group (Char. 34)	0.8004	0.4930

The dwarf mistletoes are usually quite host-specific, so hosts have been traditionally used as a major taxonomic character. In our studies, we have used principal host as a character and found that it is indeed very useful because it is highly correlated with the other characters used. Similarly, flowering group was found to be a very useful character although it has not been previously utilized in the classification of Arceuthobium (Wiens 1968).

IV. Classification

In a very strict sense, a classification for a group of organisms is a division of the group into non-overlapping, exhaustive subgroups, in order to preserve information about the organisms. A classification into species would ideally preserve information about the potential for gene exchange and interbreeding; classifications at higher levels would preserve information about the evolutionary history of the species earlier determined. Frequently the evidence sufficient to decide what the natural species for a group should be is not available. Similarly, there may be no evidence suggesting probable evolutionary trends for the species. When this happens it still can be useful to construct classifications.

Genetic and phylogenetic information are not the only kinds of information which classifications can preserve. We have seen that characters themselves contain information. Further, the information in one character can be used to predict information in other characters as well. If two characters share a great deal of information; i.e., their R-value is high, then knowledge of one provides us with some knowledge of the other. Characters are the

means we have chosen to preserve the information with which we wish to describe the collection of objects under study. In default of evidence supporting speciation or phylogenetic history, a classification may still find justification in the preservation of the information in these characters.

A classification for a group of objects can itself be thought of as a character for that group, the classes of the classification being the states of the classification character. Thus, it is possible to associate with a classification the amount of information it contains. Now , the quality of this information may be objectively measured assuming the quality of the other descriptive characters is good.

Let us agree, in the present case, that the purpose of a classification is to preserve information about the descriptive characters. A classification meets this purpose if we can predict some of the information in the descriptive characters on the basis of the information in the classification. The extent to which a classification satisfies this purpose can be measured. Let C represent a classification character and X represent a descriptive character, then C preserves information in X if H(X/C) is low, R(X,C) is high, D(X,C) is low and so forth. The information in a classification is of high quality if it can be used to predict the information in the descriptive characters. In this way the information in the descriptive characters can be summarized and preserved in a classification of high quality.

Therefore , the amount of a character's information which is preserved by a classification, C, may be measured as R(C,X). Characters

for which R(C,X) is low are those not well preserved by the classification while characters for which R(C,X) is high are easily predicted from the classification. If a classification is based on all the descriptive characters for a study, those characters which are unrelated to all the rest will be poorly represented by the classification while those related to the rest will be well preserved by the classification. Recall characters 14 and 16 in example earlier. These had highest overall distance and lowest overall redundancy. As we would expect these characters were largely ignored by the classification derived for this group (Hawksworth and Wiens, in prep.).

Some objective numerical measures of the quality of information preserved in a classification might be the sum of D(C,X) taken over all descriptive characters, where low values indicate good quality, or the sum of R(C,X) taken over all descriptors X where high values indicate good quality. There are other measures as well.

Such objective measures as those suggested above provide a means of deciding which of two competitive classifications better ~~serves to~~ preserves the information in a given collection of descriptive characters.

A comparison between the classification (Char. 35) and all characters used showed that flowering group (Char. 34) is best preserved by the classification. It has the highest redundancy (1.485) and the lowest distance (0.277) of any of the 34 characters used.

The following conditional probability table shows that the classification may be used to predict flowering group (char. 34)

TABLE 5

CLASSIFICATION (Char. 35)

(Insert Table)

For example, if a specimen is coded in flowering group state 1, there is a great chance that it will fall in Group III of the classification. There is a slight chance that it will be in Groups I or II, but no chance that it will be in Groups IV, V, or VI. If a specimen is coded in flowering group state 2, it is certain that it will fall into Group III of the classification. If a specimen is coded flowering group state 3, it is almost certain that it will fall into Group V of the classification, a slight possibility that it will fall into Group VI, and no chance of it being in Groups I, II, III, or IV. If a specimen is coded into flowering group state 4, it is certain to be in Group IV of the classification, and no chance that it will fall in Groups I, II, III, V, or VI.

V. Conclusions

i) Indication of "goodness" of characters.

Taxonomists have long recognized that characters vary considerably in their classificatory value (Davis and Heywood, 1936), but just how "good" or how "bad" a given character is for a particular set of objects has been a very nebulous consideration. In this program, each character is compared with the resulting classification, and a quantitative evaluation of each is obtained. For example, in Arceuthobium flowering group was shown to be the best of the 34 characters used because it is best preserved by the classification resulting from analyses of all characters. Conversely, anther location was the "poorest" character, and all other characters had intermediate values.

This method can be used to discover how much each character contributes to taxonomic structure by determining the quantity of information it shares with other characters.

ii) Inter-relationships of Characters

These analyses are useful for showing the relationship of a particular character to all others in the study. If, as is in the case of Arceuthobium, some characters share much common information, then they have high predictive value. For example, if the flowering state is known for a particular specimen, its probability of correct placement in the appropriate sexual dimorphism class is great.

iii) Diagnostic Keys

It has been shown that the classification can be incorporated into the analysis as another "character." By reversing this process, it is possible to identify the characters whose information is sufficient for identifying members of a particular class. Thus the characters that could be efficient in developing diagnostic keys are shown. Examples of these in Arceuthobium would be flowering group and principal hosts.

iv) Evaluation of a Classification

It is possible to measure a classification to see how well it preserves the information assimilated from all characters used in the group. Classifications which preserve much information would be more successful than those which preserve little.

v. Summary. Characters of the genus Arceuthobium (Viscaceae) are used as an example of an information theory model for character analysis. The analyses are useful in defining the quantity of information in each character, to show the relationships of characters to one another, and to show how each character contributes to the resulting classification. The method has many potential applications, not only in the biological sciences, but in other fields as well.

LITERATURE CITED

Davis, P. H. and V. H. Heywood. 1963. Principles of Angiosperm Taxonomy. Princeton, New Jersey, Van Nostrand.

Engelmann, G. 1850. Plantae Lindheimerianae, II. Boston J. Nat. Hist., 6: 214-215.

Estabrook, G. F. 1967. An information theory model for character analysis. Taxon, 16: 86-97.

Estabrook, G. F. and D. J. Rogers. 1966. A general method of taxonomic description for a computed similarity measure. BioScience, 16: 789-793.

Gill, L. S. 1935. Arceuthobium in the United States. Conn. Acad. Arts and Sci. Trans., 32: 111-245.

Hawksworth, F. G. and D. P. Graham. 1963. Dwarf mistletoes on spruce in the western United States. Northwest Sci., 37: 31-38.

Hawksworth, F. G. and D. Wiens. 1965. Arceuthobium in Mexico. Brittonia, 17: 213-238.

Hawksworth, F. G. and D. Wiens. A Monograph of the genus Arceuthobium. (In prep.)

Irwin, H. and D. J. Rogers. 1967. Monographic studies in Cassia (Leguminosae-Caesalpinoideae). II. A taximetric study of section Apoucouita. Mem. of the New York Bot. Gard., 16: 71-118.

Kuijt, J. 1960. The distribution of dwarf mistletoes, Arceuthobium, in California. Madroño, 15: 129-139.

Wiens, D. 1968. Chromosomal and flowering relationships in the dwarf mistletoes (Arceuthobium). Amer. J. Bot., 55: 325-334.

Wirth, M., G. F. Estabrook and D. J. Rogers. 1966. A graph theory model for systematic biology with an example for the Oncidiinae (Orchidaceae). Syst. Zool., 15: 59-69.

To be Supplied by Frank

6½

Fig 7

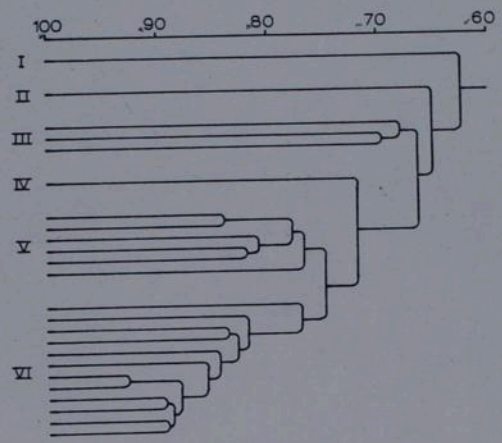


Figure 1. Classification of Arceuthobium in North America showing the division of 24 taxa into 6 groups. The scale at the top shows the relative similarity of each specimen (Estabrook and Rogers 1966).

TABLE 1*

Number of States	Frequency of Characters in this Study	Information	
		Maximum $\log_2(n)$	Range in this study
2	3	1.000	0.323 - 0.602
3	10	1.585	0.706 - 1.500
4	6	2.000	1.290 - 1.856
5	5	2.322	1.636 - 2.241
6	3	2.585	1.760 - 2.062
7	3	2.807	2.212 - 2.380
8	1	3.000	2.377
9	2	3.170	2.974 - 3.057
17	1	4.087	3.258

* Cf. pp. 3, 4, 5 and iii Examples

TABLE 2

Character	Unconditional Probability Distribution over States 1	2	3	Information
Branching type (Char. 6)	0.034	0.091	0.875	0.649
Pollen spine length (Char. 25)	0.182	0.432	0.386	1.500

TABLE 1*

Number of States	Frequency of Characters in this Study	Information	
		Maximum $\log_2(n)$	Range in this study
2	3	1.000	0.323 - 0.602
3	10	1.585	0.706 - 1.500
4	6	2.000	1.290 - 1.856
5	5	2.322	1.636 - 2.241
6	3	2.585	1.760 - 2.062
7	3	2.807	2.212 - 2.380
8	1	3.000	2.377
9	2	3.170	2.974 - 3.057
17	1	4.087	3.258

* Cf. pp. 3, 4, 5 and iii Examples

TABLE 2

Character	Unconditional Probability Distribution over States 1	2	3	Information
Branching type (Char. 6)	0.034	0.091	0.875	0.649
Pollen spine length (Char. 25)	0.182	0.432	0.386	1.500

TABLE 3

CONDITIONAL PROBABILITIES

		Sexual Dimorphism (Char. 7)	
	States	1	2
Flowering Group (Char. 34)	1	1.000	0
	2	0	1.000
	3	0.981	0.019
	4	1.000	0
Unconditional Probabilities for States of Char. 7.		0.912	0.088

TABLE 4

CONDITIONAL PROBABILITIES

		Sexual Dimorphism (Char. 7)	
	States	1	2
Shoot Color (Char. 27)	1	0.908	0.091
	2	0.933	0.067
Unconditional Probabilities for States of Char. 7.		0.912	0.088

TABLE 5

CLASSIFICATION (Char. 35)

	State	I	II	III	IV	V	VI
Flowering Group (Char. 34)	1	0.143	0.143	0.714	0	0	0
	2	0	0	1.000	0	0	0
	3	0	0	0	0	0.963	0.037
	4	0	0	0	1.000	0	0