A Computer Program for Classifying Plants

The computer is programmed to simulate the taxonomic process of comparing each case with every other case.

David J. Rogers and Taffee T. Tanimoto

A complete classification of plants is far from realization. Being well aware that plants form a dynamic, ever-changing group of organisms, no taxonomist ever expects a "complete" system of classification. Nevertheless, the possibility that a thorough, integrated system can and should eventually be produced must not be dismissed.

If no more new areas were to be explored, if no more collections of plants were to be made and stored in the world's herbaria, the number of specimens already on hand would still provide taxonomists with sufficient research material to continue their activity along present lines for at least another 50 years. An even more dismal fact is that plant taxonomists tend to neglect the plants with which we are most intimately associated, the cultivated species.

Stumbling Blocks to Progress

One of the biggest stumbling blocks to more rapid progress is the vast amount of information which must be correlated with great care, even for very small segments of the plant kingdom. Current techniques for comparison of data are largely those used by taxonomists for the past hundred years —that is, comparing, character by character, the specimens, their geographic distribution, and the ecologic and other data of various units (taxa) under study, and slowly accumulating sufficient correlation by "inspection" to allow a satisfactory system of classification. There is no doubt that the taxonomist actually, in his mind, is making comparisons of many variables when he "shuffles" his specimens, but he has been unable to convert his mental picture of these variables into a system which can be communicated readily. Actually, there is no better information retrieval system in use today than that provided by the taxonomist, zoological or botanical.

Many efforts have been made to illustrate, diagram, or reduce to formulas the various pieces of information which are found in plant classification. The most significant of these is the work of Anderson (*1*), whose beautifully simple diagrammatic techniques have been proved valuable in elucidating many problems relating to reduction of masses of data to understandable and workable units. Many efforts have been made to use statistical analyses, but we feel that many systematists have been misled when they have attempted to employ statistical formulas, which in essence measure linear phenomena, for the classification of objects with multitudes of associated variables.

Several recent efforts have been made to correlate a very large number of characteristics (*2–4*). The procedure is to give each of the characteristics equal weight; to derive correlation coefficients and determine the intervals between the correlation coefficients (*4*) which are significant for differentiation of taxa; and finally to set up a classification based on these findings. All these systems have elements which are commendable [the accompanying statements of these workers imply that the systems are closely correlated with "older" orthodox classifications of the same group (*3, 4*)]. If these methods were not so extremely laborious, they might have some chance of ultimate success, but most have been carried to such lengths that they make the process of classification even more difficult than it has been in the past. Further, if analyses were made of the numerous characters utilized in elucidation of an "unweighted" system, it would be found that within a large number of given characters, only certain sets or combinations of characters are actually significant in accomplishing the job of classification. The additional characters may be mere accessories. However, the value of correlating accessory characters with characters of high taxonomic significance, for purposes other than taxonomic, is not to be disregarded.

If one examines the taxonomic methodology, it may be easily seen that in the process employed, even in "classical" taxonomy, numbers of characters are examined, at first unweighted. Then, by a process of elimination, those not useful are dropped. "Weighting" is not a useless procedure in the process of classification. It came to be so considered among certain groups because the taxonomist took the procedure for granted and seldom used specific methods to demonstrate the importance of one character or another. The taxonomist did not feel that it was necessary to demonstrate how a character was weighted. For example, roses have compound leaves; so do some maples. This character is valueless for differentiation of species of roses, all of which have compound leaves, but of value for differentiation of box-elder maple (compound leaves) from other maple species (simple leaves). It has, therefore, a higher order of value or weight for differentiation among the species of maple than it does in the genus *Rosa.*

A Middle Way

If the problems are as outlined above, is there any middle way by which the taxonomist may speed up the process of classification? The classical methods are slow, and statistical methods, frequently too laborious, often do no more than demonstrate the obvious. Can the techniques employed be telescoped? Can the data be gathered, digested, and weighted in some way to assist the botanist and others involved in the problem of data correlation to do their job more rapidly?

Dr. Rogers is curator of economic botany at the New York Botanical Garden, New York. Dr. Tanimoto is a mathematician in the Research Division of International Business Machines Corporation, Yorktown Heights, N.Y.

The answer, we think is yes—by the application of the electronic computer to the taxonomic method. The only qualifying comment one should make is that this will not reduce the need for the trained taxonomist. The specialist must know his plants and the discipline of taxonomy as well. It would be foolish to expect a taxonomist to walk into the computer room and start pushing switches intelligently. It would be equally foolish to expect a mathematician to walk into the herbarium and start classifying plants.

If we evaluate normal procedures employed by the taxonomist—the sorting of specimens, the examination of the literature to see how earlier workers have classified the same or similar plant groups, the integration of geographic distribution, the accumulation of genetic and ecological data—and by means of the computer speed up the difficult and slow parts of the process, then the techniques described below may have some value. The program described here simulates the endless process of comparing each case with every other case in order to establish the degree of similarity between the two and thus to sort the cases into natural groupings or clusters for classification. No set judgment of the nature of a species or other taxon is implied. The taxonomic rank to be assigned to the group or cluster must rest finally with the taxonomist.

The computer program has been designed so that the investigator has full control over the process at every stage of comparing and classifying the information. Constant feedback from machine to man permits the fullest play of the subjective judgment of the botanist or other experienced person who understands the implications of the information being processed. The program is flexible enough to permit the investigator to test intuitive impressions or "hunches." It is possible at any time to eliminate a case or an attribute found to be irrelevant and to introduce new data into the system under study whenever the process of classifying indicates new directions which should be explored.

Some species, genera, and perhaps even families are satisfactorily handled by normal taxonomic procedure. Classifications of these by means of a computer would be an obvious waste of time and money. However, for complex taxonomic problems, of which a very large number exist, the computer methods seem ideally suited.

Premachine Operations

It may be helpful to those not familiar with the taxonomist's method of data gathering if we describe it briefly here.

The material used as a case study in the experimental taxonomy computer program consists of 300 herbarium specimens of *Manihot esculenta* (cassava, yuca, manioc, or tapioca) collected by one of us (D. J. R.) (*5*) in Jamaica and Costa Rica. Most specimens represent a separate cultivar, but some duplication exists. Each specimen was selected from a plot in which at least 15 plants of each cultivar were raised under conditions that were nearly identical for the 15 plants. Thus, each specimen is a rough "population sample," taken to represent, as nearly as possible, all the plants of that cultivar.

A habit photograph was made of each specimen at the time of harvesting, at a standard distance against a scaled board, and field notes were made of outstanding characteristics. The form and included information were finally standardized, and with the resulting data sheet it was possible to go directly from field data to punched card.

Choosing the Field Data

The information placed on the standardized form was chosen with specific reference to its usefulness in differentiating the cultivars. Because *Manihot* plants are vegetatively reproduced in culture and because the root produces the food, it is desirable to distinguish the cultivars by vegetative characters, and it is only of secondary importance to know the characteristics of the flower, fruit, and seed. Indeed, we have never observed some of the cultivars in flower, and they may have been so highly selected by those who use the plants for their basic food that flowering seldom occurs.

Additional data taken from the herbarium specimens were added to the standardized form after it was returned to the laboratory. All the data related to gross morphology or to pigmentation. No anatomical or biochemical information was used. Several attempts were made to include findings on the hydrogen cyanide content of the roots and findings from crude protein analyses of the foliage (*6*), but because the sample numbers were confused, it was impossible to correlate these data with all the rest.

Table 1. Characteristics and attributes.

Characteristic	Attribute	Presence
Color of stem	Gray	No
	Brown	Yes
	Yellow	No
	Red	No
Leaf shape	Linear	No
	Obovoid	Yes
	Pandurate	No

Table 2. Characteristics and attributes.

Characteristic	Attribute	Presence
Soil type on which plant grows	Acid sand	No
	Basic clay	No
	Clay and/or sand	Yes
	Volcanic soils	No
	No soil preference	No

Because most cultivars of *Manihot esculenta* occur in South America, particularly in Brazil, no final conclusions on the over-all classification of this variable species are possible. However, studies of the specimens so far collected permit a rough classification.

The computer program has been designed to analyze both qualitative data (that is, data expressed more accurately in yes-or-no form than in terms of numerical values arbitrarily assigned to describe the same qualities) and quantitative data. As an example of qualitative data, with this program it is possible to classify information dealing with "intangible" qualities such as variations in colors of plants.

The first step in preparing data is to determine which characteristics may be distinguishing within the total group and for each such characteristic to select a suitable set of attributes covering the required range of variation. Some of the main principles are illustrated in Table 1.

Clearly, the attributes must be so chosen that they are mutually exclusive (there must be no more than one *yes* per characteristic). Sometimes this re-

Table 3. Characteristics and attributes.

Characteristic	Attribute	Presence
Leaf pubescence	Glabrous	No
	Glaucous	No
	Puberulent	Yes
	Villose	No
	Tomentose	No

Table 4. Characteristics and attributes.

Characteristic	Attribute	Presence
Leaf lobe length	5–10	No
	11–17	No
	18–20	Yes

quires that a combination of attributes be listed as itself an attribute, as in a species which occurs in more than one type of soil (Table 2). To record a *yes* for the attribute "no soil preference," which means that any soil will do for the plant, is quite different from recording neither a *yes* nor a *no* for any attribute in the list, which means simply that no information is available on that attribute and, therefore, that this particular characteristic cannot be evaluated.

Many subtle characteristics will demand the subjective judgment of the botanist as to whether or not an attribute is present. For example, keen observation would be required to score correctly a characteristic such as the pubescence of a leaf (Table 3).

Some information in any system is most correctly stated in numerical form, especially measurement (of leaf length, leaf width, stamen numbers, chromosome counts, and so on). To include such characteristics in the computer processing, it is only necessary to divide or partition the range of values into a suitable set of intervals, which are then designated as separate attributes (Table 4).

Computer Processing

When all the information has been recorded in standardized form for each case, specimen, or sample, it is then punched into I.B.M. cards for computer processing. Within the computer each *yes* is recorded as the value of *one*, while each *no* is recorded as the value of *zero*. In this form, the computer can automatically compare each case with every other case, as shown in Table 5.

Table 5. Case comparisons.

Characteristic	A	B	C	D	E
Attribute	a b c d	a b	a b c d e	a b c d	a b c
Case No. 1	0 0 1 0	0 1	0 1 0 0 0	0 1 0 0	0 0 0
Case No. 2	0 1 0 0	0 1	0 0 1 0 0	0 1 0 0	1 0 0
Attributes in common in cases Nos. 1 and 2	0 0 0 0	0 1	0 0 0 0 0	0 1 0 0	0 0 0
Total distinct attributes in cases Nos. 1 and 2	0 1 1 0	0 1	0 1 1 0 0	0 1 0 0	1 0 0

Next, the computer counts the number of one's common to both and also counts the total number of distinct attributes (ones) possessed by case 1 and case 2 to arrive at the similarity ratio s_{12} for cases 1 and 2; that is, s_{12} is the ratio of the number of attributes in common in cases 1 and 2 to the number of distinct attributes possessed by cases 1 and 2 (*7*).

If no attributes for the two cases match, the ratio will be zero. If s_{12} is not zero, it will always be less than or equal to one and signifies that case 1 is related to case 2. The ratio s_{12} is one when all of the attributes match; in particular, the similarity ratio of any case with itself is always one. The ratio may be interpreted as the probability of selecting at random an attribute common to both cases 1 and 2 out of all the attributes possessed by them collectively. Hence, a similarity ratio which is nearly one would indicate a great degree of similarity between the two cases. Similarly, a small but nonzero value of the ratio corresponding to a pair of cases would indicate a nonrandom divergence of characteristics.

One may define a "distance" d_{ij} between case *i* and case *j* by

$$d_{ij} = -\log_2 s_{ij}$$

where s_{ij} is the similarity ratio corresponding to cases *i* and *j*, so that if the cases are considered as geometrical points in a suitable space, the symmetric matrix (d_{ij}) is quite analogous to a mileage chart giving the distance between case *i* and case *j*. Note that if case *i* is very similar to case *j*—that is, if s_{ij} is nearly one, the distance d_{ij} is small, so that case *i* is very close to case *j* in a geometrical sense as well as in a qualitative sense. Function d_{ij} defines what is known in mathematics as a semimetric space as contrasted with a metric space, since the sum of two sides of a triangle is not necessarily greater than the third. Designation of a space as semimetric implies that there may be two cases both related to a third in such a way that the first two may not necessarily even be related to each other. However, when a distance function defines a metric space, two cases which are both related to a third will be necessarily directly related to each other. Thus, a semimetric space seems to be more advantageous for our purposes. In terms of information theory (*8*), d_{ij} is the total information in "bits" conveyed by the event of selecting at random an attribute common to case *i* and case *j* out of the totality of distinct attributes possessed by cases *i* and *j*.

First, assuming all of the distances d_{ij} to be finite, we define an over-all number H_i associated with case *i* by

$$H_i = \Sigma_j d_{ij} = \Sigma_j -\log_2 s_{ij}$$

That is, H_i is the sum of all of the *n*-1 distances (if we have *n* cases) from the case *i* to all of the other related cases. That case i_0 for which the corresponding value of H_{i_0} is the least,

$$H_{i_0} = \min_i H_i = \min_i \Sigma_j -\log_2 s_{ij}$$

we define as the typical case. Geometrically it is obvious that case i_0, when cases are considered as points, is the point nearest the centroid of the system of points. (This point may not necessarily be unique in certain symmetrical situations.) In terms of probability theory, from the additive-multiplicative property of the logarithmic function, case i_0 is that case which is most likely to have attributes possessed by all of the others; or, in statistical nomenclature, case i_0 is determined by a "maximum likelihood" criterion. From an information-theory point of view, case i_0 possesses the least pairwise over-all total information. (That such a case i_0 is typical has been borne out empirically by a computer program in several applications.)

Before the above over-all analysis is made, the computer places each case in the order of the number of other cases to which it is related. This is done for each case by making a count R_i of the number of other cases with which case *i* has at least one attribute in common. Suppose that 100 cases are being analyzed, that case 1 has corresponding to it a value $R_1 = 95$, and that all of the other R_i values are less than 95; then case 1 can be considered as being more typical than the others. Thus, for the general situation, using the values R_i and H_i, the computer can rank all of the 100 cases for typicality, first according to their corresponding values R_i of general typicality within the entire collection and then according to their H_i values relative to the other cases having the same R_i

Table 6. "Profile" of typicality.

R value	*H* value	Case No.	Ranking
97	229.0	7	1
97	238.2	68	2
96	221.1	42	3
95	205.9	23	4
95	211.3	88	5
95	213.8	1	6
	etc.		

values. The result is a kind of profile of typicality for the collection as a whole (Table 6).

The case i_0 which ranks first in this typicality table is designated as the prime node, and around it is formed a cluster or a clump of cases very similar to it—that is, cases j, $j = 1,2,\ldots,n$, for which $1 - s_{i_0 j}$ is small.

But how similar to the prime node should another case be in order to qualify for membership in the clump? In order to determine what constitutes a clump, we introduce the notion of a measure of inhomogeneity of a set of cases. We shall say that a set of points form a single uniform clump if the distances between every pair of points are all equal; and if, in addition, the common distance is large, then each point individually can be considered a clump. This is the ideal situation: when all of the similarity ratios corresponding to distinct pairs of the n cases are equal to each other, so that the total entropy $\mathcal{E}_n$ over the system of n points defining $(n/2)(n-1)$ segments can be defined to be

$$\mathcal{E}_n = \log_2 (n/2)(n-1)$$

$\mathcal{E}_n$ is the maximum value that can be attained by an entropy function associated with $(n/2)(n-1)$ segments where the probability of selecting a particular segment is the equally likely geometric probability. In a similar fashion we define the total entropy $E_n[(d_{ij})]$ of a given set of points determined by the cases whose distances are the elements of the matrix (d_{ij}) by

$$E_n[(d_{ij})] = -\tfrac{1}{2} \sum_{ij}{}' \frac{d_{ij}}{T_n\,[(d_{ij})]} \log_2 \frac{d_{ij}}{T_n\,[(d_{ij})]}$$

where the normalization factor is given by

$$T_n[(d_{ij})] = \tfrac{1}{2} (\sum_{ij}{}' d_{ij})$$

where Σ' indicates summation only of the finite terms after repeated rows and columns are deleted. If g is the number of zeros in the symmetric matrix (d_{ij}) which lie strictly above the main diagonal, and h is the number of infinite elements above the main diagonal which are not on the same rows and columns as the g zeros, then the maximum entropy expression $\mathcal{E}_n$ must be modified because of degeneracy (that is, points coinciding) and, in addition, the lack of h segments, so that it becomes

$$\mathcal{E}_n(g,h) = \log_2 \left[\frac{n-g}{2}(n-g-1) - h\right]$$

Thus, a reasonable measure of inhomogeneity $u_n[(d_{ij})]$ determined by the matrix (d_{ij}) can be given by the normalized difference

$$u_n\,[(d_{ij})] = \frac{\mathcal{E}_n(g,h) - E_n[(d_{ij})]}{\mathcal{E}_n(g,h)} = 1 - \frac{En[(d_{ij})]}{\mathcal{E}n(g,h)}$$

This expression is identical with Shannon's definition of redundancy (*8*).

At the outset the computer can determine that a homogeneous set of cases constitutes only one clump, or that each individual case is a clump. If u_n is large, that is, near 1), we then, by means of the analysis given above, determine the ranking of typicality.

The computer now considers the cases j, $j = 1, 2, \ldots, n$, for which the distance $d_{i_0 j}$ from the prime node is less than the distance $d_{i_0 j_0}$ from the prime node, to the case j_0 which ranks second in the typicality ranking. If there are k cases within this "open" sphere whose center is the prime node, the computer determines the measure of inhomogeneity u_k of this subset; if u_k is small, the case j_0 which is on the periphery of the sphere is added to the set, and the new inhomogeneity measure u_{k+1} is computed. This process is iterated until the inhomogeneity measure suddenly takes a large jump in value, thus giving us a subset of cases constituting a clump. If u_k is large, the case nearest the periphery of the "open" sphere is removed and the measure of inhomogeneity is recomputed. This process is iterated if necessary. In this case, the worst possible situation is that we have only two members in the first clump, which may be two distinct clumps of a single case each. The prime clump determined in this way is then removed from the collection as a special class to be carefully studied. The machine can also automatically provide an analysis of the attributes in this prime clump, which would be in effect a profile of the differentiating characteristics for the central group of cases.

Similarly, an analysis can be made of attributes for the entire collection, forming a prime cluster of these attributes in the same manner as a prime cluster of cases is selected. This prime cluster of attributes constitutes a profile of the most typical attributes when all cases are considered. It can be compared with the profile of attributes for the special class of cases constituting the prime cluster, to see what significant relationships or divergencies may be revealed.

Likewise, a study can be made of the cases which rank lowest in the typicality scale, to see if a cluster exists there. In the sense of a negation, these least typical cases may reveal important tendencies which might otherwise go unnoticed among the most typical cases at the top of the scale.

Thorough study of all these first results may also lead to valuable conclusions about the usefulness or significance of the information in the original collection, with indications as to which data convey the most information for classification.

In general, prime clusters, being the most typical instances, give the least information about the branching tendencies of natural subgroups within the collection. The next step is to remove these most typical cases or characteristics, or both, from the collection. Then the remainder of the information can be reanalyzed to reveal more specialized clusters of cases and their attendant attributes. This process may be repeated, like the branching of a family tree, until the collection is narrowed down to a residue of only the most *atypical* cases. For example, these atypical cases might represent the nearest approach to the parent species in a hybrid swarm.

At any stage of the processing, the necessary decisions on what direction to take next and whether to terminate the effort can be made by those most able to judge the usefulness of the results up to that moment.

It is clear that this general method of classification, which we might call taxonometrics, can be applied to many areas where many of the data are qualitative (*9*).

References and Notes

1. E. Anderson, *Introgressive Hybridization* (Wiley, New York, 1949).
2. P. H. A. Sneath and S. T. Cowan, *J. Gen. Microbiol.* **19**, 551 (1958).
3. C. D. Michener and R. R. Sokal, *Evolution* **11**, 130 (1957).
4. J. A. Soria, thesis, Indiana Univ. (1958).
5. This work was supported in part by grant No. 2327 from the National Science Foundation.
6. D. J. Rogers, *Econ. Botany* **13**, 261 (1959).
7. The similarity ratio is the ratio of the dimensions of elements of a Boolian lattice. See G. Birkoff, "Lattice Theory," vol. 25 of *American Mathematical Society Colloquium Publications* (rev. ed., New York, 1948). Also see J. von Neumann, "Continuous Geometry," *Proc. Natl. Acad. Sci. U.S.* **22**, 92 (1936).
8. C. E. Shannon, *Bell System Tech. J.* **27**, 379, 623 (1948).
9. The donation of staff and computer time to this study by the International Business Machines Corp. is gratefully acknowledged.