information connected with a concept. For example, for a given polymer its method of preparation and further processing, its properties, uses, etc., can be brought together at the node allotted to that polymer in the graph. Thus, inspection of the information allotted to that node is sufficient to ascertain whether or not the information of interest for that polymer is indeed presented in that document. The document can then be very quickly accepted or rejected as a response to an inquiry.

Once a chemist has learned to read the graph for a document, he will advantageously make use of this method for representing and thereby illustrating complicated concept connections he encounters in his profession (and in his daily life!). An example is phrasing a patent claim of one's own or comparing different patent claims with respect to their degree of overlap or the gaps existing between them.

There is every indication that such a guide to the graphical representation of concept connections is sufficiently simple, efficient, and general to be employed on a versatile basis.

## ACKNOWLEDGMENT

This publication is dedicated to Professor Dr. Schultheis, the initiator of IDC, on the occasion of his 70th birthday.

## LITERATURE CITED

(1) Fugmann, R., Nickelsen, H., Nickelsen, I., and Winter, J. H., "TOSAR—A Topological Method for the Representation of Synthetic and Analytical Relations of Concepts," *Angew. Chem., Int. Ed. Engl.*, **9**, 589–595 (1970).

# A Comparison of the Performance of Some Similarity and Dissimilarity Measures in the Automatic Classification of Chemical Structures

GEORGE W. ADAMSON* and JUDITH A. BUSH

Postgraduate School of Librarianship and Information Science, University of Sheffield, Western Bank, Sheffield S10 2TN, England

Received October 14, 1974

**A group of 39 structures with local anesthetic activity has been classified automatically by calculating similarity or dissimilarity coefficients between pairs of structure diagrams and applying cluster analysis to the results. The performance of a number of similarity and dissimilarity coefficients has been compared using the relationship between structure and property. Simple coefficients and a distance function give more satisfactory results than functions using probabilistic weighting or standardized distance.**

Techniques for the automatic classification of chemical structures could have application in the storage and retrieval of chemical information[1] and in pattern recognition studies on chemical data.[1,2]

The 20 common naturally occurring amino acids have been classified by Sneath using numerical taxonomic methods, and on the basis of a manual analysis of their structures and some of their physical, chemical, and biological properties.[1,7] More recently the same compounds were classified using automatic procedures and solely on the basis of their structure diagrams.[1] As the structure diagrams of chemical compounds can be directly related to their properties,[2] then it is possible that structural parameters derived from structure diagrams will be useful in pattern recognition calculations on chemical data.[1,12]

The classification method used in the work described below is broadly similar to that applied to the amino acids;[1] however, it is applied to a group of 39 structures with local anesthetic activity.[6] The local anesthetics are structurally more diverse than the amino acids and thus illustrate the effectiveness of the method when applied to a heterogeneous set of structures.

The classification was carried out using different measures of structural relationship, and their performance was compared by using the measures of relationship and the classifications derived from them to simulate the prediction of the log (MBC), *i.e.*, minimum blocking concentration values of the compounds. The performance of the similarity (SC) and dissimilarity (DC) coefficients is thus estimated in a way which would be useful in situations where the relationship between the structure and the property is important.

## METHOD OF CLASSIFICATION

The structure of each anesthetic was described as a redundant connection table, and this was used to obtain a set of augmented atom fragments[5] upon which measures of association were based. The same fragment type was used in the classification of 20 naturally occurring amino acids[1] and consists of an atom, the bonds formed by the atom, and the atoms to which it is bonded, excluding bonds to hydrogen atoms. Single and double bonds in rings and chains were also differentiated in this investigation.[1]

The anesthetics were first analyzed to identify the different augmented atoms occurring, and based on these a set of attributes was chosen to represent each structure. The following two descriptions were used.

(i) For each augmented atom type identified, a suitable set of attributes was selected to cover the different occurrences in each structure. Thus each attribute in the given set was used to indicate whether or not the particular fragment type was present in a structure at the given frequency. Using this qualitative description, multiple occurrences of the same fragment in a structure were then accounted for by additive coding.[19]

(ii) A single attribute was chosen to represent each augmented atom type and it indicated the number of occurrences in a structure of the given fragment type.

* Author to whom inquiries should be addressed.

In case (i) a binary vector was set up to describe each structure, and in case (ii) a vector whose attribute values corresponded to augmented atom frequencies. The SC or DC between each pair of structures was then calculated from the corresponding pair of vectors.

The first three coefficients considered relied on a two-state attribute description[17,19] and were applied to structure representation (i). For each pair of structures the data were arranged in the form of a 2 × 2 table in which attributes were categorized into four groups a, b, c, and d, where $a$ is the number of attributes which are common to both structures, $b$ and $c$ are the numbers which occur in the first structure but not the second and *vice versa*, and $d$ is the number which occurs in the set of structures but in neither of the pair of structures under consideration. The three coefficients were

$$\text{Dice's SC} = \frac{2a}{2a + b + c} \tag{1}$$

$$\phi = \frac{(ad - bc)}{[(a + b)(a + c)(d + b)(d + c)]^{1/2}} \tag{2}$$

$$\text{Sneath's DC} = \frac{b + c}{a + b + c + d} \tag{3}$$

The numbering of expressions in the text corresponds to the numbers used in Table I. These coefficients were also used to classify the naturally occurring amino acids[1] when the best results were obtained using structure representation (i). This representation was one of a number based on two-state attribute descriptions which were used.

Various measures of association have been proposed in numerical taxonomy for dealing with a quantitative description of the data. One of these is the distance coefficient where attribute values are assumed to be metric quantities, and the similarity between members of a group of objects is given as a function of their distance in an $n$-dimensional space, whose coordinates are based on the attribute set used to define the objects. A number of distance measures have been suggested[7,13,18] and some of these have been used in applications of pattern recognition techniques to chemical data.[12] Because of their application in this area it seemed appropriate to consider such an association measure in the present investigation. The distance measure considered in this case was based on structure representation (ii). The occurrence of each augmented atom type in a structure was regarded as a metric, and the similarity between pairs of structures was expressed in terms of their distance apart in an $n$-dimensional space where the axes correspond to the $n$ augmented atom types defining the structures. The squared distance was used as a measure of dissimilarity and calculated using the expression below

$$\delta_{j,k}^2 = \sum_{i=1}^{n} (X_{ij} - X_{ik})^2 \tag{4}$$

where $X_{ij}$ is the number of occurrences in the $j$th structure of the augmented atom fragment defined by attribute $i$. Because of differences in the units of measurement, and in the range of values for each attribute, scaling is sometimes necessary to obtain reasonable results. In the present case, to compensate for the variation in the range of frequency of occurrences of augmented atom types, distances were recalculated after first standardizing values so that each attribute possessed a zero mean and a unit variance (coefficient 4a). In all of the association measures considered so far, equal importance has been given to each fragment type considered.

The weighting of characteristics used in the classification process is an issue in numerical taxonomy over which there seems to be no general agreement at present. Many arguments have been put forward for weighting.[11,19] One of these is that infrequently occurring attribute states are more discriminating and should thus be more heavily weighted than frequently occurring states. Several coefficients have been proposed based on this criterion,[8-10,14-16] and some of these were used in the classification of the anesthetics.

In the comparison of a pair of structures the weight attached to a particular attribute pair was calculated from the likelihood of that pair of states or a more similar pair of states arising.

In all the coefficients considered in order to reduce the amount of computation the weighting process was based on the approximation that different attributes in a structure occurred independently of each other.[4] In the comparison of structures, attributes were considered in turn and the weight attached to the particular pair of values arising for each attribute was based on Goodall's definition of similarity for individual attributes.[8] An overall measure of similarity between each pair of structures was then obtained by summing the similarity terms over all attributes, using the following expression

$$\sum_{i=1}^{n} - \log P_{ijk}$$

where $P_{ijk}$ is the similarity term derived for attribute $i$ in structures $j$ and $k$. $n$ is the total number of attributes. The definition of similarity depended on the type of attribute in question. The first probability coefficient considered (5) was based on structure representation (i), and Goodall's definition of similarity for qualitative attributes. The second probability coefficient (6) was based on structure representation (ii) and used Goodall's definition of similarity for metrical attributes. The third probability coefficient considered (7) was also based on structure representation (ii), except that in this case the definition of similarity for each attribute was based on attribute values alone. The SC's or DC's obtained using the above coefficients were used to form clusters by the single linkage method.[11,19] The cluster analysis was carried out using an agglomerative algorithm developed by van Rijsbergen.[20]

## RESULTS

**a. Similarity and Dissimilarity Coefficients.** The performance of the similarity and dissimilarity coefficients was compared by simulating their use to predict log (MBC) values. Agin, Hersh, and Holtzman[6] considered the relationship between minimum blocking concentration and other properties of the compounds. They derived an expression relating log (MBC) with the polarizability and ionization potential of the molecules and obtained a good correlation which could be used for predicting the local anesthetic activity of compounds. The observed values for log (MBC) given by Agin, *et al.*, are used in the results described below.

In order to obtain a predicted value for the log (MBC) of a particular structure, the log (MBC) value of the anesthetic with which it was most closely associated was used. Where more than one nearest neighbor arose, the average log (MBC) value over the set of nearest neighbors was used. For each association measure the sum of the squares of the difference between observed and predicted log (MBC) values, taken as a ratio of the sum of the squares of the deviations of the observed values from their mean, was calculated. The average value of the difference between observed and predicted log (MBC) values was also calculated. These were used as an indication of the effectiveness of the measure of association. The results are shown in Table I. The best result was obtained using the squared distance coefficient, where the sum of squares ratio was 0.34 and the mean deviation between observed and estimated values was 0.79. To put these results into perspective, a comparison was made with the mean deviation of the observed values from their mean, the best possible result which could be expected for the given set of values, and the mean deviation which would have resulted if there had been no

**Table I.** Log (MBC) Estimations for a Group of 39 Local Anesthetics, Based on a Number of Different Measures of Association and the Classifications Obtained Using These[a]

| Measure of association | | Predictions based on highest SC or lowest DC | | Predictions based on classification | |
|---|---|---|---|---|---|
| | | $\frac{\sum_{i=1}^{N}(x_i-\hat{x}_i)^2}{\sum_{i=1}^{N}(x_i-\bar{x})^2}$ | $\frac{\sum_{i=1}^{N}\lvert x_i-\hat{x}_i\rvert}{N}$ | $\frac{\sum_{i=1}^{N}(x_i-\hat{x}_i)^2}{\sum_{i=1}^{N}(x_i-\bar{x})^2}$ | $\frac{\sum_{i=1}^{N}\lvert x_i-\hat{x}_i\rvert}{N}$ |
| 1 | | 0.527 | 0.994 | 0.543 | 1.167 |
| 2 | | 0.662 | 1.071 | 0.664 | 1.273 |
| 3 | | 0.428 | 0.843 | 0.819 | 1.300 |
| 4 | | 0.343 | 0.786 | 0.659 | 1.076 |
| 4 | (a) | 0.611 | 1.069 | 0.859 | 1.429 |
| 5 | | 0.928 | 1.454 | 0.950 | 1.579 |
| 6 | | 1.618 | 1.891 | 1.420 | 1.973 |
| 7 | | 0.516 | 1.000 | 0.679 | 1.207 |

[a] $x_i$ is the observed property value, $\hat{x}_i$ the predicted property value, $\bar{x}$ the mean observed property value for the group, and $N$ the total number of structures in the group. The numbering of the coefficients is the same as that used in the text.

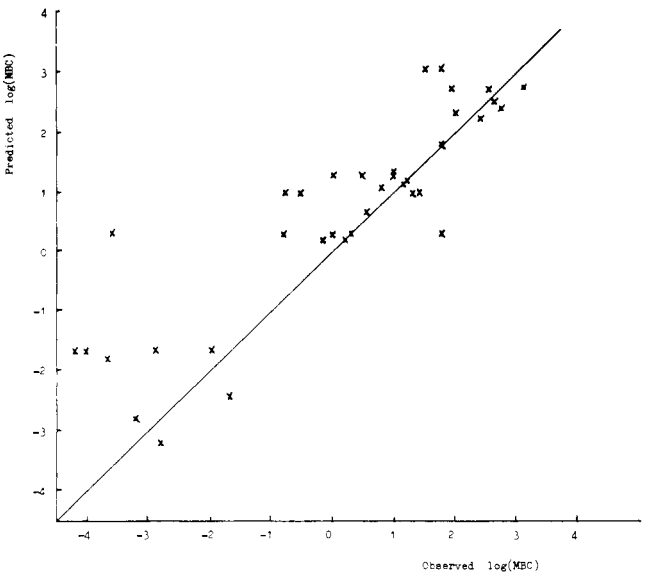


**Figure 1.** Graph of observed against predicted log (MBC) values using the simple distance coefficient (coefficient 4) together with the 45° line on which all points would lie if predictions were perfectly accurate.

resolution of the structures into classes. These values are 1.65, 0.09, and 1.69, respectively. The second value would be obtained if each anesthetic had its highest SC or lowest DC with the anesthetic or anesthetics which also had the closest log (MBC) value. In the case of the third value, the predicted property for each anesthetic would be the average log (MBC) value for the remaining 38 structures in the set. The mean deviation of 0.79 using the squared distance coefficient is therefore an improvement on both the mean deviation for the set and the mean deviation assuming a homogeneous group. The worst result was obtained using coefficient 6, the probability coefficient based on quantitative attributes, where the mean deviation between observed and estimated log (MBC) values exceeded both these values. The extent of the relationship between the property in question and the structural differences as measured by the distance coefficient is shown in Figure 1, which gives a plot of observed log (MBC) values against the values predicted on the basis of this DC.

A number of the anesthetics are very well predicted, *e.g.*, quinoline, 8-hydroxyquinoline, and acetanilide, and included in these is a group of normal alcohols ranging from ethanol to 1-octanol. Methanol is not included in this group owing to the absence in its structure of a methylene group. This makes its association with ethanol very much weaker than the latter's association with propanol. It is interesting to note in the case of the alcohols that each member of the group, except terminal members, is equally highly associated with the two alcohols adjacent to it in the series, and therefore the predicted property value is exactly that which would be obtained by linear interpolation from the nearest neighbors in the homologous series. Some structures in the group which did not belong to such a distinct chemical class and which formed no other strong associations, were poorly predicted, *e.g.*, quinine and eserine.

**b. The Classifications.** The classifications were similarly assessed on the basis of their predictive power. The predicted property for each anesthetic was taken to be the average log (MBC) value for the cluster which it joined. The results are shown in Table I. Dice's coefficient gave the lowest sum of squares ratio, and the mean deviation between observed and predicted log (MBC) values was again lowest using the squared distance coefficient. The worst predictions were obtained using coefficients 5 and 6, both of which took the probability of occurrence of fragment types into account in assessing similarity. In the latter case, where structure representation (ii) was considered, the mean deviation of 1.891 was again greater than the sample mean deviation and the mean deviation resulting if no resolution of structures into classes had taken place. Except for coefficient 6 the association measures gave a better prediction than the classifications. This result is not unreasonable in view of the information loss accompanying the transformation from an association matrix to a dendrogram and the diverse structural types represented. The dendrogram illustrating the classification based on the squared distance coefficient is shown in Figure 2. Within this heterogenous sample some fairly distinct clusters have been formed. The majority of structures show a very early breakdown into cyclic and acyclic classes, with the exception of antipyrene which is associated with the acyclic group and of eserine, diphenhydramine, phenyltoloxamine, caramiphen, and quinine which have cyclic and acyclic components of comparable size. These form no strong associations with the remainder of the set and join classes by chaining at a much lower level of similarity. The smaller group of acyclic structures reveals a well-defined cluster of *n*-alcohols. Except for the terminal members of the group each alcohol is equally highly associated with adjacent members with a DC value of 1, resulting in the formation of a cluster at this level. Methanol and 2-propanol join this cluster at a slightly lower level owing in both cases to an association with ethanol.

The group of cyclic structures shows first a general breakdown according to the size of ring substituent, with structures incorporating larger acyclic components separating from those without this feature. At a higher level of similarity the former group tends to cluster according to the nature of the ring substituent, whereas the latter breaks up according to the nature of the ring system.

The latter group reveals a well-defined cluster of simple benzene derivatives consisting of toluene, phenol, benzyl alcohol, and hydroquinone. Quinoline and 8-hydroxyquinoline also form a cluster at this level. This second group does not include the remaining N-heterocycles present due in most cases to the influence of the acyclic component. In the case of benzimidazole, however, it is because the five-membered heterocyclic ring was coded with localized ring bonds. For the group containing larger acyclic substituents, one main cluster is formed, but at a much lower level of similarity, between procaine, tetracaine, and xylocaine, two of which are dialkylaminoethyl esters of benzoic acid and the other a dialkylamino derivative of acetanilide. Dibucaine and caramiphen which have very similar acyclic substituents are not closely associated with this group owing to

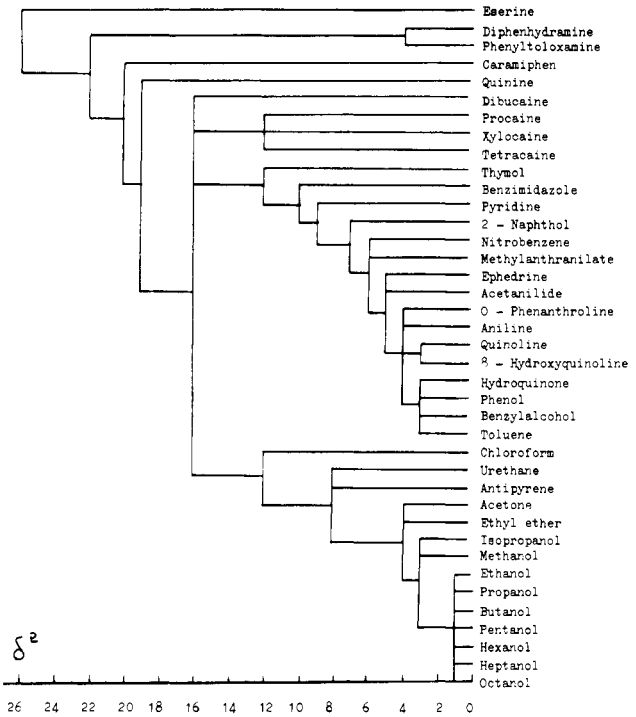


**Figure 2.** Dendrogram illustrating the classification obtained using the above distance coefficient and the single-linkage clustering method.

the influence of the ring systems present. This also applies to the structural isomers phenyltoloxamine and diphenhydramine which are dissociated from the two main groups of cyclic and acyclic structures, but which are themselves closely associated and form a separate cluster at a DC value of 4.

## CONCLUSIONS

A hierarchic classification of 39 structures with local anesthetic activity has been derived by an automatic method. The classification produced is consistent with a qualitative view of the relationships between the structures. The performance of the similarity and dissimilarity measures and of the classifications in predicting log (MBC) values shows that similarity in structure as determined by the methods used is paralelled by a similarity in property.

The simple coefficients using binary attributes and additive coding and the distance functions based on quantitative attributes perform better in this case than the functions involving probabilistic weighting. It will be interesting to see if this result is also obtained with other properties and groups of structures.

In view of the possibility of applying pattern recognition methods to prediction it is worthwhile comparing the best prediction in this work, mean deviation 0.786 and sum of squares ratio 0.343, with other methods. A mean deviation of 0.18 and a sum of squares ratio of 0.01 was obtained from the results of Agin, *et al.*[6] A regression analysis based on augmented pair occurrence produced a mean deviation of 0.07 and a sum of squares ratios < 0.01.[3] Thus, if quantitative prediction of properties based on structure diagrams is the main objective, then pattern recognition methods are not the only methods which should be considered. The relationship between structure and property which is produced by the classification and SC's and DC's indicates that these techniques could usefully be incorporated in information storage and retrieval systems.

## EXPERIMENTAL DETAILS

Computer programs were written in Plan (the ICL assembly language), Fortran, and Algol, and were run on the Sheffield University ICL 1907 computer, which has a 24-bit word length and a cycle time of approximately 2 μsec. The analysis of connection tables and description of structures in terms of augmented atoms was carried out by a Plan program which also incorporated Plan and Fortran subroutines for the calculation of SC's and DC's (CPU times ≤14 sec, core storage used ≤2785 words + working storage). A Plan program incorporating van Rijsbergen's clustering algorithm as a Fortran subroutine was used to cluster the structures, after first arranging the SC's and DC's in decreasing order of similarity or increasing order of dissimilarity. Each SC or DC value is examined and a listing of the clusters formed at each level of association is produced, except where these are identical with those formed at the previous level (CPU times ≤9 sec, core storage used 3672 words + working storage). The programs used for property prediction were written in Algol.

## ACKNOWLEDGMENTS

We wish to thank Drs. M. F. Lynch and U. D. Naik for useful discussions and the Office for Scientific and Technical Information (London) for the award of a Postgraduate Research Studentship to J. A. Bush.

## LITERATURE CITED

(1) Adamson, G. W., and Bush, J. A., "A Method for the Automatic Classification of Chemical Structures," *Inform. Stor. Retr.*, **9**, 561–568 (1973).
(2) Adamson, G. W., and Bush, J. A., "Method for Relating the Structure and Properties of Chemical Compounds," *Nature (London)*, **248**, 406–407 (1974).
(3) Adamson, G. W., and Bush, J. A., unpublished results
(4) Adamson, G. W., Lambourne, D. R., and Lynch, M. F., "Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-Based File. Part III. Statistical Association of Fragment Incidence," *J. Chem. Soc., Perkin Trans. 1*, 2428–2433 (1972).
(5) Adamson, G. W., Lynch, M. F., and Town, W. G., "Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-Based File. Part II. Atom Centered Fragments," *J. Chem. Soc. C*, 3702–3706 (1971).
(6) Agin, D., Hersh, L., and Holtzman, D., "The Action of Anaesthetics on Excitable Membranes: A Quantum Chemical Analysis," *Proc. Nat. Acad. Sci., U.S.*, **53**, 952–958 (1965).
(7) Bielecki, T., "Some Possibilities for Estimating Inter-population Relationships on the Basis of Continuous Traits," *Current Anthrop.*, **3**, 368; discussion, 20–46 (1962).
(8) Goodall, D. W., "A New Similarity Index Based on Probability," *Biometrics*, **22**, 882–907 (1966).
(9) Goodall, D. W., "A Probabilistic Similarity Index," *Nature (London)*, **203**, 1098 (1964).
(10) Harrison, P. J., "A Method of Cluster Analysis and Some Applications," *J. Appl. Stat.*, **17**, 226–236 (1968).
(11) Jardine, N., and Sibson, R., "Mathematical Taxonomy," Wiley, London, 1971.
(12) Kowalski, B. R., and Bender, C. F., "Pattern Recognition. I. A Powerful Approach to Interpreting Chemical Data," *J. Amer. Chem. Soc.*, **94**, 5632–5639 (1972).
(13) Penrose, L. S., "Distance, Size and Shape," *Ann. Eugen.*, **18**, 337–343 (1954).
(14) Rogers, D. J., and Tanimoto, T. T., "A Computer Program for Classifying Plants," *Science*, **132**, 1115–1118 (1960).
(15) Smirnov, E. S., "On the Expression of Taxonomic Affinity," *Zh. Obshch. Biol.*, **27**, 191–195 (1966).
(16) Smirnov, E. S., "On Exact Methods in Systematics," *Syst. Zool.*, **17**, 1–13 (1968).
(17) Sneath, P. H. A., "Relations between Chemical Structure and Biological Activity in Peptides," *J. Theoret. Biol.*, **12**, 157–195 (1966).
(18) Sokal, R. R., "Distance as a Measure of Taxonomic Similarity," *Syst. Zool.*, **10**, 70–79 (1961).
(19) Sokal, R. R., and Sneath, P. H. A., "Principles of Numerical Taxonomy," W. S. Freeman, San Francisco, Calif., 1963.
(20) van Rijsbergen, C. J., "A Fast Hierarchic Clustering Algorithm," *Comput. J.*, **13**, 324–326 (1970).