

AUTOMATIC CLASSIFICATION AND  
CHEMICAL STRUCTURE - ACTIVITY CORRELATION

A Thesis submitted for the  
Degree of Doctor of Philosophy

at

The University of Sheffield

by

J. A. Bush

Postgraduate School of Librarianship and  
Information Science

February 1977

### ACKNOWLEDGEMENTS

I would like to express much gratitude to my supervisor, Dr. G. W. Adamson for his continual advice and guidance throughout the course of this study; and to Professors M. F. Lynch and W. L. Saunders for their help and encouragement, and for making the investigations possible.

I also wish to thank Dr. U. D. Naik for valuable discussions on statistical techniques, and the Department of Education and Science (London) for financial support during the period of research.

## SUMMARY

A review of the available literature on automatic classification methods in chemical structure applications has shown that there has been surprisingly little interest in the use of unsupervised approaches in this area, considering the potential value of these in large structure-based chemical information systems. In the first section of the thesis the suitability of such an approach, using a simple hierarchical clustering technique and an approximate structure representation based on fragment sets derived from the structure diagram, was investigated. Using a connection table representation of structures, the feasibility of combining the classification method with automatic substructure handling techniques important in current computer-based systems was also considered. Finally, the value of results based exclusively on two-dimensional substructural descriptors was assessed.

Preliminary studies using a simple binary representation of structures and recording only the fragments shared between each structure pair demonstrated the feasibility of the approach. Detailed studies were then carried out to compare alternative methods of structure representation and comparison, the former involving investigations of both substructural descriptors and their numerical representation. Generally accepted evaluation procedures were not available to test the success of methods and the classifications and association measures used in their derivation were assessed partly on chemical significance and partly on predictive performance.

More detailed numerical representations, based on the number of occurrences of the fragment types in a structure, gave better separations of structures and also better predictions than representations based only on the presence or absence of a fragment. In the former case better

results were given by definitions which distinguished between equivalent substructures occurring in chains and non-aromatic ring systems. In the comparison of structures, simple matching coefficients and a simple Euclidean distance measure performed as well as more complicated measures using fragment weighting, and the simpler coefficients often gave a better result, both in structure separation and predictive performance. Also, the coefficients based on quantitative fragment descriptions were no better than those based on simple binary representations using additive coding. The use of standardised characters with the distance function gave poor results. Coefficients showing the best separation of structures gave best predictions, but prediction levels were close and differences were difficult to interpret. Similar trends, however, were observed in a number of different samples suggesting that the results are of some significance. In contrast, different substructural definitions did not perform similarly in different samples, and in two small data sets, one involving similar structural types and the other very dissimilar types, opposite trends were observed in predictive performance. Another larger related group showed little variation. There was some within sample consistency between classification and predictive performances although the fluctuations shown in the two small samples were not paralleled by equally wide variations in structural arrangements and the significance of these prediction results would need to be tested further. Atom descriptions gave poor predictions and classifications in most cases. However, in the small structurally diverse sample they gave a good prediction due to the particular distribution of functional groups in this sample i.e. the occurrence of groups important for activity in dissimilar structural types with similar molecular formulae. This result therefore was not considered particularly significant and other

structurally diverse groups involving different structure-property relationships are not expected to behave in this way. The wide disparity between classification and predictive performances in this example, however, illustrated the practical difficulties involved in choosing suitable methods and showed how this could depend on the particular application.

The above investigation clearly demonstrated the potential of an unsupervised classification approach for structuring large data bases and dealing with both closely related and diverse structural types. The good agreement between observed and 'predicted' property data in this work also suggested the method could be useful in structure-property correlation studies. This was investigated in the second section of the thesis by comparing the approach with an alternative empirical method based on regression analysis. The analyses were carried out under similar conditions to the classifications, and like the classification approach the regression model developed is the first of its kind able to look at structure-property relationships in diverse sets of structures, and to use automatic procedures of substructural analysis.

Structure-property agreement in the regression case did not vary widely with the size of substructures although larger fragments gave lower residual errors, and in some cases a more significant correlation. Furthermore, the use of higher order relationships did not lead to a significant improvement over a linear function. An assessment of predictive performance using predictions simulated by the 'hold-one-out' technique showed that this was not simply related to the significance of the correlation. However, the regression coefficients required for prediction were not always available and with more suitable 'learning sets' the more significant correlations are expected to give a better result. Interpretation of the regression

solutions was limited both by the approximate nature of substructural definitions and their interdependency. Nevertheless, many of the coefficients were statistically significant and although coefficients themselves did not differ significantly each substructures contribution to the property in question could be explained sensibly in physicochemical terms, giving good agreement with the results obtained in other similar investigations. The regression solutions therefore had potential value in rationalising structure - property relationships, and in biological applications they could aid more detailed analyses.

The classification and regression methods gave similar levels of prediction, although under equivalent conditions ie. using the same substructural definitions, the regression equations always gave the better result. This suggested some difference between approaches, which would not be unreasonable in view of the more accurate nature of the regression method. Comparisons with other similar structure - property studies based on pattern recognition and additive statistical modelling, showed that both methods were potentially useful for quantitative prediction. Correlations in the regression case were also as successful as those obtained in semiempirical studies, using quantum - chemical or linear free energy related parameters to describe structures. Additionally, the two approaches dealt equally well with diverse structural groups and could be used in early drug design studies to investigate possible new leads.

CONTENTS

	PAGE
INTRODUCTION .. .. .	1
Chapter 1 NUMERICAL CLASSIFICATION .. .. .	8
1.1 Background .. .. .	8
1.2 The Basic Approach .. .. .	9
1.3 Unsupervised Learning Methods .. .. .	10
1.4 General Advantages of a Numerical Approach to Classification .. .. .	12
1.5 Numerical Classification based on Cluster Analysis .. .. .	13
1.5.1 Defintion of Terms .. .. .	13
1.5.2 Defining the Data .. .. .	16
1.5.2(a) Choice of Characters .. .. .	16
(i) Conceptual Problems .. .. .	16
Nature of Classifications	16
Nature of Classes .. .. .	18
(ii) Statistical Problems .. .. .	20
1.5.2(b) Importance of Characters .. .. .	21
(i) 'A prior' and 'A posteriori' weighting .. .. .	21
(ii) Character Probabilities .. .. .	22
(iii) Correlation and Redundancy .. .. .	24
1.5.2(c) Choice of Numerical Representation .. .. .	27
(i) Qualitative values .. .. .	27
(ii) Ordered values .. .. .	28
(iii) Quantitative values .. .. .	28
(iv) Missing values .. .. .	31

	PAGE
1.5.3 Estimation of Resemblance .. ..	31
1.5.3(a) Association coefficients ..	32
1.5.3(b) Distance coefficients ..	34
1.5.3(c) Probabilistic coefficients ..	37
1.5.3(d) Correlation coefficients ..	39
1.5.3(e) Choice of Resemblance Measure	40
1.5.4 Methods of Cluster Analysis .. ..	42
1.5.4(a) Clustering Methods and Applications .. ..	43
1.5.4(b) Choice of Clustering Method ..	48
1.5.5 Evaluation Problems .. ..	50
 Chapter 2 THE STRUCTURE DIAGRAM IN CHEMICAL INFORMATION HANDLING AND STRUCTURE-PROPERTY CORRELATION .. ..	 52
2.1 Introduction.. .. .	52
2.2 The Structure Diagram .. .. .	53
2.2.1 Use in Chemical Communication ..	53
2.2.2 Structure-Property Relationships ..	54
2.2.3 Use in Structure-Property Studies	55
2.2.4 Comparison of Structural and Physico- chemical Parameters in Structure- Property Studies .. .. .	58
2.2.5 Choosing Suitable Substructures ..	60
2.3 Automatic Classification Methods .. ..	63
2.3.1 Supervised Learning .. ..	63
2.3.2 Visual Display .. ..	66
2.3.3 Unsupervised Learning .. ..	67
2.3.4 Comparison of Supervised and Unsuper- vised Learning Approaches .. ..	68
2.3.5 A Novel Classification Method for Handling Chemical Structures .. ..	69



	PAGE
2.4 Regression Analysis .. .. .	70
2.4.1 The Semi-empirical Model .. .. .	70
2.4.2 The Empirical Model .. .. .	71
2.4.3 A Novel Empirical Regression Model based on Explicit Structure Definitions .. .. .	73
2.5 A Comparison of Regression Analysis and Pattern Recognition in Quantitative Structure-Property Studies .. .. .	74
2.5.1 Data Requirements .. .. .	75
2.5.1(a) The Dependent Variable .. .. .	75
2.5.1(b) The Independent Variables .. .. .	76
2.5.2 Evaluation Procedures and Available Statistical Criteria .. .. .	77
2.5.2(a) Regressions .. .. .	77
2.5.2(b) Classifications .. .. .	79
2.5.3 Property Prediction .. .. .	80
2.5.4 Computational Considerations .. .. .	81
 Chapter 3 A METHOD FOR THE AUTOMATIC CLASSIFICATION OF CHEMICAL STRUCTURES .. .. .	 84
3.1 Introduction .. .. .	84
3.1.1 The Basic Approach .. .. .	84
3.1.2 Evaluation Problems .. .. .	85
3.1.3 Feasibility Study .. .. .	86
3.1.4 Data Sets .. .. .	87
3.2 A Comparison of Some Alternative Numerical Representations of Substructures .. .. .	88
3.2.1 Method .. .. .	89
3.2.1(a) Substructures .. .. .	89
3.2.1(b) Numerical Representations .. .. .	89
3.2.1(c) Structure Comparison .. .. .	90

	PAGE
3.2.1(d) Clustering .. .. .	91
3.2.2 Results and Discussion .. .. .	92
3.2.2(a) Predictive Performance ..	92
(i) The Similarity and Dissimilarity Coefficients	93
(ii) The Classifications ..	94
3.2.2(b) Structural Arrangements ..	95
3.3 An Evaluation of Some Different Measures of Resemblance .. .. .	97
Introduction .. .. .	97
3.3.1 Method .. .. .	100
3.3.1(a) The Structure Representations	100
3.3.1(b) The Similarity and Dissimilarity Coefficients ..	101
(i) Association/Matching Coefficients ..	101
(ii) Distance Coefficients ..	102
(iii) Probabilistic Coefficients	103
3.3.2 Coefficient Performance .. .. .	107
3.3.2(a) Nearest Neighbours .. ..	107
3.3.2(b) Classifications .. ..	111
(i) Simulated Predictions ..	111
(ii) Structural Arrangements ..	113
3.4 Choosing Suitable Substructures .. .. .	118
Introduction.. .. .	118
3.4.1 Method .. .. .	119
3.4.1(a) The Association Measure ..	119
3.4.1(b) The Structure Representation ..	120
3.4.1(c) The Fragments .. .. .	120

	PAGE
3.4.2 Fragment Performances .. ..	121
3.4.2(a) Prediction Levels .. ..	121
(i) Nearest Neighbours ..	121
(ii) Classifications .. ..	122
3.4.2(b) Structural Arrangements ..	124
3.5 Conclusions .. .. .	136
 Chapter 4 THE DEVELOPMENT OF AN EMPIRICAL STRUCTURE-PROPERTY CORRELATION METHOD BASED ON REGRESSION ANALYSIS ..	142
4.1 Introduction .. .. .	142
4.2 The Empirical Model .. .. .	145
4.3 Method .. .. .	146
4.3.1 The Input Matrix .. .. .	146
4.3.2 Dependent Variables .. .. .	146
4.3.3 Independent Variables .. .. .	147
4.4 The Correlations .. .. .	148
4.4.1 Structure-Property Agreement .. .. .	148
4.4.2 Use of the Correlations for Property Prediction .. .. .	151
4.4.3 Interpretation of the Regression Solutions .. .. .	155
4.5 Conclusions and Comparisons with other Regression Approaches on the Same Data .. ..	158
 Chapter 5 DISCUSSION OF THE CLASSIFICATION AND REGRESSION APPROACHES AS METHODS FOR STRUCTURE-PROPERTY CORRELATION .. .. .	163
DESCRIPTION OF COMPUTER PROGRAMS .. .. .	167
APPENDICES .. .. .	172
I Data Sets and Properties .. .. .	172
II Altering the Sample to Feature Ratio in Classification Applications .. .. .	190
III Semiempirical Structure-Property Correlation using Structural Parameters .. ..	192

BIBLIOGRAPHY	..	..	..	..	..	..	198
--------------	----	----	----	----	----	----	-----

Tables

Figures

Publications

## INTRODUCTION

In recent years there has been considerable interest in numerical classification techniques in research areas of information science concerned with the development of more efficient data handling techniques. Most of the research has been directed towards improving retrieval strategies for large document collections.<sup>1-3</sup> But increasingly the method is becoming important for a variety of data analysis problems e.g. property estimation, which in the past have been dealt with by techniques such as factor analysis, principal component analysis and regression analysis.

Despite the widespread interest in numerical classification methods for handling bibliographic data<sup>1-3</sup> there has been very little application of the approach to chemical structure information. The purpose of the present study is to see whether suitable methods can be developed in this area. With the growing interest in automatic procedures for the design of new drugs<sup>4-9</sup> the study considers the value of automatic classification for property prediction as well as for structure retrieval. In the former case its suitability is evaluated by comparing structure-property correlations with those given by a new empirical method based on regression analysis.

There is now a wide range of automatic classification techniques available and the different approaches are discussed in detail in Chapter I. The particular approach considered in this investigation is based on cluster analysis, where structures are grouped according to the relationship between individual members of the group under consideration. These relationships must first be expressed quantitatively and in turn the statistical measures of association used to obtain them require that the structures first be represented in numerical form. There are therefore a number of different stages involved in the classification process, each of which requires approximations which will affect the final result.

One of the major problems in automatic classification is to define

meaningful numerical representations of the original data. The accuracy of the numerical descriptors depends on a number of factors, such as the nature of the original data and the type of association measure considered. In the case of a chemical structure a variety of representations are possible, some of which provide a more accurate description of the real structure than others. The structure diagram, which is the level of structural description used throughout the present investigation, is only an approximate two-dimensional projection of the real structure, but it is an important starting point because of its widespread use in chemical information systems and in chemical communications in general.

The literature shows that in the few applications where chemical data have been subjected to automatic cluster-based classification procedures a combination of structure and property data is usually considered.<sup>10,11</sup> In addition, structural descriptions are usually chosen on the basis of their assumed diagnostic importance.<sup>10-13</sup> The main objective of the present investigation is to devise methods for handling the structural attributes of chemical species using techniques which could be easily incorporated in existing computer-based chemical information systems, and which could be applied automatically.

The structure diagram may be represented in a variety of ways for the purposes of computer manipulation, but for explicit and unambiguous definitions connection tables or linear notations are usually employed.<sup>14-16</sup> The methods developed here are based on a connection table representation. This is broken down automatically into sets of substructures which are suitable for setting up the appropriate numerical representations for structure comparison. Because of the simplicity of the connection table record and its very close relationship to the structure diagram, the fragmentation process is straightforward and the algorithm developed is fast and simple. Algorithms of this type had been developed previously for use

in computer-based substructure search systems,<sup>17</sup> in which similar numerical representations are considered. The substructures obtained do not uniquely define the structure diagram, and incorporate some redundancy, the extent of which depends on the size of the substructure being used. Finally, additional approximations must be made in setting up numerical records which can be used as a basis for structure comparison. The extent of the approximations in this case depends on which association measure is used and the type of numerical descriptors which are appropriate for its application. The difficulties arising in obtaining meaningful comparisons between individuals are discussed in detail in Chapter 1, and the specific problems arising in the case of chemical structures are discussed in Chapters 2 and 3. Often the choice of representation is restricted by the type of association measured considered and vice versa. Both qualitative and quantitative numerical representations have been considered here, and a variety of association measures capable of handling these, ranging from simple matching coefficients to distance measures and probabilistic similarity functions. Probabilistic measures, unlike distance and simple matching coefficients have not been extensively applied, although a wide variety of such measures have been proposed. This is because of the large amounts of computation usually involved in calculating them. The probability measures considered in this investigation had previously not been applied and thus particular attention is paid to their performance compared with that of the non-probabilistic measures.

Finally the individual estimates of resemblance between structures must be summarised in a way which will reveal meaningful chemical groups. Again, a wide range of methods is available at this stage, but in this case the difficulties arising in choosing appropriate methods is largely independent of the nature of the original data. Because of this, the study has concentrated on evaluating different structural representations and estimates of resemblance between structures, and has used for this purpose

a simple hierarchic clustering technique throughout.<sup>18</sup> This and other similar clustering techniques are described in Chapter I. Variations in the first two stages of the classification process have been considered separately so that their effect could be clearly assessed, and full details of the methods developed are given in Chapter 3.

In choosing suitable methods cluster evaluation constitutes a major problem because of the absence of widely accepted evaluation procedures. This is partly because of general disagreement over the objectives of classifications and partly due to the mathematical properties of the method. The problems involved and the methods of evaluation currently in use are discussed in Chapter I. In many applications classifications are expected to have inductive properties and this has become an important criterion for judging classification performance.

The method of evaluation considered here is based on the assumption that the structural features of chemical compounds are related to their physical, chemical and biological properties. However, because the structure diagram is only an approximate representation of the real structure it provides only a limited basis for the prediction of the properties of the molecule it describes. The expected imperfect correlations between structure and property data nevertheless provides a basis for the comparison of the methods developed. The classifications and association measures used to derive them were compared by simulating the prediction of an observable property in each case, and determining the extent of the agreement between observed and predicted property values. Whether or not the classifications and association measures are suitable tools for predictions however depends on the approximations in the method. Thus, in order to estimate their predictive value it was necessary to compare the predictions with some carried out by alternative approaches.



Quantitative property estimations in chemical structure applications are more usually based on structure-property correlations using regression analysis.<sup>4-9</sup> This approach was considered a suitable alternative here as it is a more exact approach with widely accepted procedures of evaluation. The regression analysis methods developed to date relate property data either empirically to a set of structural features, or semi-empirically to known physicochemical parameters, which in turn are related to structure. Usually methods are concerned with variations in side chain structures, and property data is related to these only. But in this study the whole molecular structure was taken into account to explain the property in question. This new approach increases the usefulness of the structure-property correlation method, as it enables a wide range of structural types to be examined simultaneously. The important consequence of this is that the method can be used to explore possible new lead structures,<sup>5</sup> in contrast with existing methods which are aimed at optimising activity within a given chemical series. The new regression approach is discussed in detail in Chapter 4. Basically the property of the set of structures under consideration, taken as the dependent variable is assumed to be related linearly or by some other simple function to the structural attributes of the compounds which are expressed as a set of independent variables. Provided the correlations obtained are significant they are then used as a basis for prediction. The usual tests of significance were applied and were used to compare the suitability of a number of different structural representations. Predictions were simulated using the 'hold one out' technique.<sup>5,19,20</sup> Each structure in turn is removed from the set of structures under investigation and a property value is estimated for it from the results of the regression analysis on the remaining structures in the set. Details of the methods developed are given in Chapters 2 and 4.

Because of the potential application of this new empirical approach to structure-property correlation the methods developed here were considered as a possible tool for property prediction, in addition to providing a basis for the evaluation of the classification work, and where possible the results were compared with other regression approaches described in the literature. Comparisons however were made difficult for the reason that very few other investigations reported to date have tried to use the correlations obtained by regression analysis for property prediction. Comparisons with the classification work were also limited because the large numbers of substructural fragments necessary to describe whole structures often prevented a regression analysis.

The classification and regression methods were tested using a number of small data samples extracted from the literature. From the results obtained the suitability of methods for larger scale applications is considered in view of the computational difficulties expected. It is possible in such small investigations as this that the results may be influenced by the failure of the sample to adequately represent the population<sup>2,21</sup> and this is taken into account both during the comparison of methods and in the consideration of larger scale applications.

The regression analysis and pattern recognition techniques described could be of value in a wide range of applications. The classification approach for example could be put to numerous uses in chemical information systems. Thus, file structures based on this technique, which brings together chemically similar structures could lead to more effective structure retrieval strategies and could also be used to obtain specialised sub-files from general data bases. The method could also be used for classifying substructure search output. One of the stages involved in the classification process is the calculation of similarity or dissimilarity

coefficients between the structures to be classified, and if suitable coefficients could be developed then these could be used to rank search output in order of their relevance to the search question. Depending on the type of coefficient used relevance could be measured on an ordinal or even more precise scale. In addition to employing the classification technique for file organisation and manipulation purposes it may also be possible to use the relationships derived between structures to bring out relationships between structure and property data. This would considerably increase the usefulness of the approach in chemical information systems in which properties and structure diagrams are already available in machine-readable form.<sup>22-24</sup> The new regression method described could also be of considerable value in this reas. It is the first statistical correlation technique developed which can handle diverse structural types. This may lead to a better understanding of the contributions to activity of different substructural features, which could in turn increase the value of the approach as a diagnostic tool in drug design. With suitable data the method could therefore be used to explore possible new lead structures<sup>5</sup> in addition to providing a useful empirical tool for property prediction. Where applicable it is expected that the regression methods developed will be a better method for prediction than the classification approach, and will therefore be the preferred approach in applications where quantitative structure-property correlation for property prediction is the main objective. Depending on the type of application and the type of data available both approaches could be valuable in computer-based chemical information systems based on the structure diagram. The scope and limitations of the two approaches are discussed more fully in Chapter 2.

CHAPTER 1

Numerical Classification

Background

Much of the early work on numerical classification was carried out in the biological sciences, where it is usually referred to as numerical or mathematical taxonomy. As early as 1898 Heincke<sup>25</sup> used a phenetic distance measure to distinguish between races of herring and in 1909 Czekanowski<sup>26</sup> employed a distance coefficient in physical anthropology. One of the first statistics extensively applied was the "Coefficient of Racial Likeness" developed by Pearson<sup>27</sup> in 1926, although this has been considered mainly by anthropologists and has not been taken up by taxonomists in general. This measure is a type of similarity coefficient and it was ultimately developed by Mahalanobis into a "Generalised Distance" statistic.<sup>28</sup> Other similar statistics were developed by Anderson and Abbe<sup>29</sup> and Anderson and Whitaker<sup>30</sup>. The growth of automatic methods was slow initially, and most of the early statistics were used mainly as discriminant functions to help identify new individuals and place them in existing classification schemes. They were therefore of limited use, and did not lead to any major advances.

Following this initial work in the natural sciences the use of numerical classification methods spread gradually to other areas, although until more recently the main application outside the natural sciences was concentrated in the behavioural sciences. Here some early applications are those by Zubin<sup>31</sup> in 1938 and Thorndike in 1953.<sup>32</sup> One of the main difficulties impeding progress in the early years was the lack of adequate processing facilities, and it is only within the last decade or so, with the

general availability of automatic computing facilities to take on the burden of the large amounts of computation usually involved, that the use of numerical classification methods has become widespread. One of the most important advances in the natural sciences was the application of cluster analysis and these methods opened the way to present-day numerical classification techniques. Work in this area was initiated by Sneath<sup>33,34</sup> Michener and Sokal<sup>35</sup> and Sokal and Michener<sup>36</sup> in the late fifties. Clustering techniques have now been successfully applied in many different areas, and there has been a great proliferation of methods in the last few years. Attempts to categorise these and produce comprehensive reviews however has been difficult because of the diverse nature of applications. Reviews of methods and applications are usually directed towards a particular subject area. Possibly because of the longstanding application of numerical classification techniques in the biological sciences a particularly wide range of literature is available in this area, and some very useful reviews have appeared, such as those by Johnson<sup>37</sup>, Blackwelder<sup>38</sup>, Sneath,<sup>39</sup> Williams and Dale<sup>40</sup>, and Sokal et al<sup>41</sup>.

## 1.2 The Basic Approach

The various approaches to automatic classification in use today are often collectively referred to as non-parametric methods of pattern recognition. Within different fields these methods have been given a variety of different titles, and some of the more common ones, such as numerical classification, automatic classification, mathematical taxonomy and numerical taxonomy, are often used interchangeably. The basic aim of these methods is to reveal the essential and otherwise unidentifiable relationships within

data sets by summarising the available information on individual members. An important characteristic which the different approaches share is that no assumptions are made about the underlying statistical distribution of the data in question.

Automatic classification procedures are of two basic types. If they are required to fit new data into existing classification schemes the classification rules employed must first classify correctly the existing information. This is often referred to as supervised learning. If the classification process is required to identify meaningful clusters in previously unknown distributions of individuals the clustering rules used are not based on available information concerning class identity, and this process is usually referred to as unsupervised learning. The present investigation is concerned mainly with applications of this second type.

### 1.3 Unsupervised Learning Methods

The majority of unsupervised classification methods begin with the calculation of the degree of resemblance between the individual members of the data sample. If the nature of the data is such that classes are very distinct, or if the sample is small then these measures may be sufficient to reveal the underlying structure of the data without the application of involved mathematical clustering procedures. Usually however such procedures are needed to bring out the essential relationships present. Two approaches have become important in recent years for this purpose. Firstly, the methods which partition the data into groups according to predefined rules on the definition of clusters and class membership. These methods are usually referred to collectively as methods of

cluster analysis. Secondly, there are the ordination or mapping techniques which summarise the available information on individual relationships so that individuals can be conveniently represented in two or three dimensions for visual display purposes. Using this second approach the individuals are initially assumed to be distributed through an n-dimensional hyperspace whose coordinates represent the features used to describe them. Some confusion has arisen over different terminologies and the terms clustering and cluster analysis are often used to encompass all the various approaches possible, including display methods. Whichever approach is used the basic objective is to summarise the relationships in the data in a way which will result in the smallest possible loss of information. However the choice of suitable method is often a difficult one, as the uncertain mathematical properties of the methods make it impossible to estimate the extent of the data loss 'a priori'. This problem is discussed in later sections of the present chapter.

Until recently ordination procedures have been less widely applied than methods of cluster analysis but they are now increasingly used, and are often applied in conjunction with clustering techniques.<sup>20, 42-45</sup> Using this approach there are several ways of reducing the data for visual display purposes.<sup>46,47</sup> Procedures such as factor analysis, principal component analysis and principal coordinate analysis have been widely considered, particularly in the behavioural sciences.<sup>48,49</sup> Multidimensional scaling techniques are also of importance, and these techniques, usually referred to as linear and non-linear mapping techniques, have recently been used in chemical structure applications to aid other pattern



recognition methods in the investigation of structure-property relationships.<sup>20</sup>

#### 1.4 General Advantages of a Numerical Approach to Classification

Numerous problems arise when applying classification methods. Some of the more general conceptual difficulties involved are discussed below in sections 1.5.2(a) and (b). Over and above these difficulties many additional problems arise when applying numerical techniques. What, therefore, can be gained from using a numerical approach. Sneath and Sokal<sup>46</sup> have recently enumerated some of the possible advantages, and those of relevance here are now discussed briefly.

Compared with conventional classification methods a numerical approach increases objectivity by reducing the number of arbitrary decisions to be made. Investigators in favour of the conventional classification approach however view this particular advantage with some doubt as they feel that arbitrary decisions based on intuitive reasonings are essential for a meaningful result. A less questionable advantage is that the approach allows much of the classification process to be automated. This is important in areas where large amounts of information are involved. Another benefit arising from this is the ability of the method to handle larger numbers of characteristics, which reduces the dangers of arbitrary pre-selection of features in the description of individuals. These were important properties influencing the initial interest in the approach in biological applications, where the expanding volume of data and the growing numbers of characteristics used to represent it were becoming increasingly difficult to handle by conventional means. Because data is

held in numerical form another advantage is that the required information for classification could be easily integrated in existing computer-based systems. In some areas automatic classification is leading to a vital revision of existing ideas, for example in the biological sciences, and in many applications the method is becoming important for its heuristic value.<sup>40, 46, 50</sup> As well as generating hypotheses the approach is also of value in shedding new light on existing hypotheses, and examples of this may be found in the behavioural sciences.<sup>51,52</sup> Finally, numerical classification has considerable potential as a tool for prediction and there have been numerous reports in the literature illustrating its possible value in this area, for example Sneath<sup>10</sup>, Kowalski and Bender<sup>12</sup>, Paykel<sup>53</sup>, Ting et al<sup>54</sup>, and Chu<sup>55</sup>.

## 1.5 Numerical Classification Based on Cluster Analysis

In most cluster-based classification applications there are three basic stages involved. Initially numerical representations of the original data must be chosen which provide a suitable basis for the comparison of individuals. Using these and a statistical measure of association, quantitative estimates of similarity or dissimilarity between individuals are then obtained. Finally, a set of clustering rules are applied to these quantitative measures held in matrix form.

### 1.5.1 Definition of Terms

The development of numerical classification techniques in a wide range of disciplines has led to an equally wide range of terminologies in defining methods. As seen earlier the classification process itself comes under a variety of different headings

depending on the field in which it is applied e.g. terms such as numerical taxonomy, mathematical taxonomy, taxometrics and systematics are usually considered in the biological sciences. Other terms such as non-parametric pattern recognition, cluster-analysis, Q-analysis, grouping, clumping and classification are used in mathematical applications, sociology, psychology and information retrieval. The terms numerical classification, automatic classification and pattern recognition are used in the present study and occasionally the labels supervised and unsupervised learning are used in cases where a distinction is being made between these different approaches.

The different terminologies arising in defining methods has added to the many conceptual problems involved in describing the classification process. This is particularly true in some of the earlier stages of the classification process, and there has been much confusion over the definition of data and the relationships arising between the original data and its representation in numerical form.

In the following discussion the members of the data set undergoing classification are referred to as entities, objects or individuals. These are broken down into a number of descriptive features referred to as characters or features. The nature of these depends on the way individuals are fragmented. Thus, a character may represent a particular aspect of the individual which is either present or absent, or else it may take on a number of separate values. These are referred to as character states, character values or characteristics. In the case of characters representing a single characteristic, characters and characteristics

become equivalent. The numerical descriptors used to represent characters are referred to as attributes, and the different values which they may take are referred to as attribute states. The nature of attributes depends both on the type of numerical representation chosen and on the nature of the characters they represent and, as with characters and characteristics, there may or may not be a 1:1 relationship between characters and attributes. Thus, characters representing single, qualitative characteristics may be represented by a single attribute, whereas multi-state characters which cannot be represented conveniently in this way must be represented by a set of attributes which cover the required range of variation. The different types of qualitative and quantitative character definitions which may arise and the possible numerical representations of these will not be discussed any further here as these are described in detail in section 1.5.2(c).

Although the above definitions have been adhered to as far as possible, it is difficult to be completely consistent in the use of these terms, particularly with such terms as character, and characteristic, which are considered to have equivalent meanings in every day usage. In the discussions preceding the description of character types the features of individuals are discussed in more general terms and no distinctions are drawn between characters and characteristics. At this stage descriptive features have been referred to as characters and occasionally the term characteristic has been used where it was felt that this was more appropriate.

### 1.5.2 Defining the Data

Data representation is a particularly critical stage of the classification process and is one which involves a number of separate issues, some of a fundamental nature. First of all the important characters of the individuals in question must be identified. Having chosen these, the relative importance of characters must be decided upon. Finally, a suitable numerical representation must be chosen which will convey the required information. The first two issues, concerning the choice and significance of characters involve more fundamental questions which arise whether or not numerical procedures are adopted, but this important point is often overlooked, and has been the cause of much unfair criticism of automatic classification methods.

#### 1.5.2.(a) Choice of Characters

##### (i) Conceptual Problems - Nature of Classification - Nature of Classes

To deal satisfactorily with the questions of character choice and character importance it is necessary to know the precise nature of classifications and their objectives. However from the time of the Greeks up to the present day there has been no universal agreement over the purposes of classifications, and as a result these properties are difficult to define.

##### Nature of Classifications

Some of the earliest ideas on systematic classification were based on Aristotelian logic.<sup>56-58</sup> However although this approach was initially widely considered it is strictly only suitable for simple, logical systems, where the individuals undergoing classification can be defined in such a way that the remainder

of their properties can be automatically inferred. Eventually these ideas gave way to a set of general principles<sup>59</sup> which were considered to be of universal applicability. These principles were based on the premise that there cannot be one ideal and absolute scheme of classification for any particular set of objects but that there must always be a number of classifications which differ according to the purpose for which they have been constructed. Many of the currently used techniques in numerical classification however have been influenced by the classification views held in the biological sciences where, until recently, these general principles have been largely ignored. The early development of taxonomic theory in this area was based on the belief that living things belong to ideal or 'natural' systems and are governed by special laws laid down by a Creator. The general principles of classification which had been widely used in the case of inanimate objects were therefore considered inappropriate and biological classification took a very different course.

The early development of taxonomic theory before Darwin was based on Lindley's concept of 'natural affinity'.<sup>60</sup> This was a very vague concept explained in terms of a 'Plan of Creation'. Following the theory of evolution, evolutionary considerations were in general thought to be essential for an understanding of natural systems, and the concept of 'natural affinity' was re-interpreted in terms of these relationships. However, evolutionary characteristics are not easy to define and in many ways this redefinition helped to widen the gap between biological and other types of classifications, as the concept of 'natural affinity' could now be even more broadly interpreted. From this time there

was very little agreement over the importance of evolutionary characteristics and their relative value in explaining natural affinity compared with observable characteristics. This controversy eventually led to general disagreement over the interpretation of natural systems and by the beginning of this century this in turn had led many investigators in the field to question the concepts upon which such systems were based.<sup>61-67</sup> The main influence in this area came from Gilmour<sup>65-67</sup> who believed that the isolation of biological classification from classification in general had been damaging and was largely responsible for much of the confusion which existed. More recently, and especially since the consideration of numerical techniques, when biologists and others were forced to re-examine their objectives, Gilmour's views have been more widely supported. However, many of the old ideas persist, and the objectives of classifications continues to be a controversial issue. In the biological sciences in particular, opinion is still very much divided. Some of the old concepts are still firmly upheld and there continues to be disagreement over the relative importance of observable and evolutionary features. Many investigators now feel that difficulties such as these will never be resolved until the principles of classification have themselves been thoroughly re-evaluated.

#### Nature of Classes

Despite differing views over classification objectives, it is generally agreed that the usefulness of classifications will depend on the number of characteristics used to define the individuals or objects in question, and that a classification which utilizes all known characteristics is more generally useful

than one based on a more limited range. However, whereas the traditionalists regard such classifications as approximations to a single, ideal classification scheme others feel that these should be considered as flexible arrangements which change as new knowledge is acquired, but which never aim towards a single end.<sup>66</sup> This latter view is in keeping with the general principles of classification and with Gilmour's basic philosophy.

One of the characteristics of the early classification schemes based on Aristotelian logic is that for class membership each individual is expected to possess all the properties which were used to define the class in question. Such arrangements are now usually referred to as monothetic groups. Most applications today however are based on polythetic arrangements where the criterion for class membership is based on the numbers of shared attributes between individuals. Polythetic classifications were first considered in the natural sciences. At the time when the concept of 'natural affinity' was first introduced and larger numbers of characteristics were involved, it was soon realised that the members of classes did not necessarily possess any one diagnostic character i.e. any one feature which is common to all class members.<sup>68</sup> This is now regarded as one of the essential characteristics of polythetic classes, although such arrangements were not formally defined as such until much later.<sup>69</sup> Recent definitions of polythetic groups distinguish two basic types. Thus, polythetic groups are defined as those whose members have a large number of characteristics in common but where no character is either essential for class membership or is sufficient to allow membership. Fully polythetic groups must satisfy the above



conditions and in addition require that no feature be common to all its members. The large numbers of descriptors usually involved in present day applications often prevent this last condition being met. Consequently most polythetic arrangements are of the former type.

From the previous discussion on classification objectives, polythetic arrangements are obviously more suitable for general purpose classifications or for classifications where the objectives are not well defined. If the objectives are more specific then monothetic arrangements may be more appropriate, although in this case there is a high risk of misclassification when the number of descriptors considered is large. Some arrangements of this type have recently been criticised by Sneath and Sokal,<sup>46</sup> e.g. Maccacaro<sup>70</sup> and Williams and Lambert.<sup>71</sup>

(ii) Statistical Problems

The availability of improved computing facilities in recent years has made it possible to consider larger numbers of characteristics during the classification process. The arguments in favour of large numbers of descriptors to obtain more general or 'natural' classifications are discussed in the previous section. Leaving these aside there are additional problems concerning the desirability of this approach from a statistical point of view. Past investigations of supervised learning methods have shown that a small sample to feature ratio can have an adverse effect on the classification result, and in the case of two-way classification schemes, for example, it has been shown that a sample to feature ratio below about 3 is undesirable.<sup>72,73</sup> However, most of the research in this area has been carried out on this type of

application and although a similar dependence on the sample to feature ratio is not expected with unsupervised learning methods it is not certain what effect, if any, the ratio has on the classification result in this case. Some investigators have stressed the need for large numbers of features in the unsupervised case to reduce the risk of distortion in estimating degrees of similarity.<sup>46</sup> However the choice of features here usually depends on the particular requirements of the user, and whether he wishes to derive a specialised classification or a more general one with wider predictive powers.

#### 1.5.2.(b) Importance of Characters

Related to the problem of character choice is the question of character significance. When numerical methods were first introduced in the biological sciences character weighting was considered an essential part of the classification process. The relative importance of different characters however was based largely on intuitive judgements and the differing interpretations of character importance led to widespread confusion. The introduction of automatic procedures was therefore seen as an ideal opportunity to revise existing ideas concerning character values, and consequently most of the numerical techniques considered at that time employed equally weighted characters. Because of this, equal weighting is often wrongly associated with numerical methods and is assumed to be an essential feature of the numerical approach.

##### (i) 'A priori' and 'A poster<sup>o</sup>iri' weighting

Many arguments have been put forward in favour of character weighting and an equally large number against it. Increasingly, forms of 'a priori' weighting, where the value of characters is estimated prior to classification, are considered unacceptable

although some who are in favour of equal weighting are in agreement with some forms of 'a posteriori' weighting.<sup>46,47</sup> However, some still criticise the arguments which have been put forward for equal weighting and claim that this process is itself a form of 'a priori' weighting.<sup>75,76</sup> Jardine and Sibson<sup>74</sup> have recently suggested that much of the disagreement over character weighting has arisen due to a failure to distinguish between forms of 'a priori' and 'a posteriori' weighting. They claim that much of the controversy concerns only 'a posteriori' weighting, as most of the so called 'a priori' arguments are usually based on some previous knowledge. An example of this is the case of expedient weighting discussed by Inglis,<sup>75</sup> where a character preselection process is applied to reduce large numbers of characters to within workable limits. This is not an instance of true 'a priori' weighting as the preselection process in such cases is usually based on previous evidence of character importance. It is generally accepted however that in most types of application some forms of character weighting, whether desirable or not, are unavoidable. The above case of expedient weighting is an example in question.

(ii) Character Probabilities

Much attention has been given to the problem of character weighting in biological applications. The weighting process here is usually based on intuitive judgements, but an issue of more general relevance which has recently been given attention in many different fields is the question of whether or not the statistical distributions of characters should also be taken into account in deciding on character importance. The

arguments put forward in favour of character weighting in this case have their basis in probability theory. Thus, should characters which arise infrequently in the set of objects or individuals in question be considered more important than frequently occurring characters? Secondly, should characters which are highly correlated with other characters be considered less important? This second question involves a number of separate issues which will be considered later.

The usual argument used in favour of weighting is that infrequently occurring characters are more discriminating, and should be weighted more heavily because of their diagnostic value. Several different weighting procedures based on character frequencies have been proposed, and where a quantitative approach is considered it is more usual for the character frequencies to be taken into account during the comparison of individuals, rather than during the preceding stage of character definition.<sup>74</sup> Thus, during the comparison of characters the likelihood of a particular pair of values arising in two individuals is determined and the less probable this co-occurrence the more similar the individuals are said to be with respect to the given character. An example of this approach is the similarity index derived by Goodall.<sup>77</sup> In defining similarity he considers character value frequencies in conjunction with the usual definition of this term, and applies these criteria to both ordered and metric data. In the case of data which is qualitative and unordered character value probabilities are the only consideration. This approach has been considered in the present investigation and the particular methods developed are discussed in Chapter 3. Other similar approaches have been proposed by Smirnov,<sup>78</sup> and Rogers and Tanimoto.<sup>79</sup>

Many workers have criticised the use of probabilistic procedures in numerical classification, as in order to apply them it is necessary to make certain assumptions about the underlying statistical distribution of the data. For example, Goodall's method is based on the null hypothesis that the values which each character may take are randomly distributed amongst the individuals in question with probabilities equal to their observed relative frequencies. Several investigators feel that this approach is unsuitable for classification purposes in the case of finite data sets e.g. Williams and Lance<sup>50</sup> and Williams and Dale<sup>40</sup>, because it is impossible to obtain null hypotheses which are independent of the given data set, and that if such procedures are employed, different samples must inevitably lead to different results. They therefore consider these procedures to be invalid. Others have rejected probabilistic methods for similar reasons.

(iii) Correlation and Redundancy

Closely related to the problem of character frequencies is the question of character correlation. A number of different problems are involved here, for which there are again no generally accepted solutions. Many different types of correlation have been discussed in the literature.<sup>46, 74</sup> All of these, regardless of their particular nature, will result in some degree of redundancy, but the main difficulty lies in determining exactly how they affect the result and whether or not these effects are desirable. In the extreme case, where one characteristic always arises in conjunction with another, and thus always implies the presence of the other, then this may be thought of as total character redundancy. However, whether this is redundancy in the true sense of the word will depend on the nature of the association.

Secondly, as in the above case of character frequencies, there is the problem of handling finite data sets and determining whether the correlations observed are simply a characteristic of the particular data sample in question. Recently in attempting to resolve these problems some investigators have expressed concern over data sampling and the need for adequate sample size.<sup>21,46,50</sup>

The differing terminologies which have arisen to explain the various types of association possible have added to the problems arising. For example, Sneath and Sokal<sup>46</sup> and Jardine and Sibson<sup>74</sup> both discuss logical correlations but their definitions do not coincide. Most agree that care must be taken in the choice of characters to avoid, where possible, true redundancy in definition, for example, where two different characters are considered one of which is simply a re-expression of the other, or where two characters are related logically and the one can be thought of as a re-expression of the other. This second example coincides with Sneath and Sokal's definition of logically redundant characters. They give the example of two characteristics one of which defines the presence of haemoglobin and the other defines the redness of blood, where the latter is dependent on the presence of haemoglobin. However, this example is straightforward but difficulties arise when the dependency is only partial. Jardine and Sibson define logically related characters as all those which are conditionally related. Thus they include in this category characters which are known to be related empirically, but are not necessarily related logically in the sense given above. Associations of this type, defined by Sneath and Sokal as empirical correlations, are sample dependent i.e. the correlation observed in one data set need not necessarily arise in any other. Jardine and Sibson have pointed out the dangers

involved in this case and suggest these may often be overcome by careful choice of characters at the outset. However, whether or not such correlations can be successfully eliminated still leaves unanswered the question of their desirability. Some investigators feel they should be accounted for in some way, and in cases where they cannot be eliminated, some form of weighting procedure should be applied.<sup>80-83</sup> Others feel that correlations should not be eliminated indiscriminately, and as all successful classifications rely on their presence it is important to distinguish between those which have an adverse effect on the result and those which are essential for the formation of sensible clusters.<sup>74</sup> Evaluating their effect however is not an easy task because of the mathematical limitations of the method and the wide range of applications, making direct comparisons of weighted and unweighted characters difficult. To give an example of the sort of problems arising Rohlf<sup>84</sup> recently investigated 45 different species of North American mosquito and concluded that character redundancy should be avoided where possible as it causes elongation of generic clusters, and makes individuals at the periphery of clusters appear more isolated than they should be. Power<sup>85</sup> on the other hand points out that if the degree of correlation varies within each of the species studied, then this information is important for discriminating between groups, and should be retained at all costs. These opposing views typify the arguments appearing in the literature over the desirability of character correlations, and unless numerical classification is placed on a more formal basis there is little chance of resolving such differences of opinion fully.

1.5.2.(c) Choice of Numerical Representation

The final step in the data preparation stage or preprocessing phase is the conversion of the chosen identifiers or descriptors into a form suitable for computation. Approximations are unavoidable at this stage and it is vital that the numerical descriptors chosen are as representative as possible of the original information, while at the same time providing a suitable basis for comparison.

The types of character definitions possible and the measurements of similarity and dissimilarity for which they are suitable have been widely discussed in the literature.<sup>46</sup> Basically two different types of information may arise, namely measurement data, usually referred to as quantitative data, and secondly qualitative data which refers to some kind of descriptive quality such as colour variation. Both types of information may be dealt with in a variety of ways.

(i) Qualitative values

Depending on the way characters are chosen, qualitative descriptions may either be represented by two-state or binary descriptors (attributes), or by qualitative multi-state descriptors (attributes). For example if a character is chosen which represents some characteristic which is either present in an individual or absent then a binary representation is used. If the character chosen may take on a number of different values which are unordered and which may or may not be linked logically, then there are various ways of representing this character numerically so that the various possibilities may be identified. One of the more usual ways is to select a suitable set of two-state attributes to cover the required range of variation. However, in the case of logically



related characteristics such as colour differences this type of coding presents a number of difficulties, as the descriptors or attributes chosen in this case are mutually exclusive i.e. the presence of one of the values, (giving a positive score on one of the attributes) automatically infers the absence of the remaining possibilities and a negative score on all remaining attributes in the set. This gives rise to two basic problems. Firstly, if two individuals do not agree with respect to this character the coding method gives rise to two mismatches. Secondly, if a similarity or dissimilarity measure is used which takes into account the agreement of negative scores then the degree of similarity is exaggerated and the extent of the distortion will depend on the number of two-state attributes employed. Thus, although this method has been used by some investigators e.g. Rogers and Tanimoto,<sup>79</sup> Sneath and Sokal<sup>46</sup> have recently suggested that the approach be used for logically independent characteristics only.

(ii) Ordered values

Where the character chosen may take on a number of different values which are not quantitative but which belong to an ordered series, a set of two-state attributes may again be employed. However, it is not the most satisfactory approach in this case, as difficulties would arise in ensuring that values closer together in the series are considered more similar than those which are further apart. This could best be accomplished by assigning arbitrary numerical values to the series and treating these in much the same way as measurement data (see below).

(iii) Quantitative values

Quantitative characters, where the different possibilities arising

are both ordered and metrical present fewer problems, and here the choice of definition often depends on the type of association measure considered. Thus the different values which a character may take may be represented by numerical quantities which coincide with the original measurement values, or, if a binary representation is required, the measurements can be broken down into a number of two-state attributes in much the same way as ordered characters described above, where suitable numerical values have been arbitrarily defined. If the values are continuous they may be divided into a suitable set of intervals, each of which is considered as a separate attribute.

However, whether the original measurement data is discrete or continuous, a few problems are presented by this breaking down process. The process may be carried out in a number of ways depending on whether the resulting two-state attributes for a given character are to be regarded as additive or non-additive.<sup>46</sup>

Where the data is continuous, and each attribute chosen must represent a range of values, the attributes are mutually exclusive and it is important that the class intervals be as small as possible to minimise the information loss. In the case of characters which take on discrete values a number of different representations are possible. The additive and non-additive coding procedures which are commonly used for qualitative characters are equally applicable here. These procedures have been discussed in detail by Sneath and Sokal<sup>46</sup> and are only described briefly below.

Using the additive coding method, if a character represents a series of numerical values ranging from 0 to  $n$ , these would be represented by  $n$  two-state attributes, all of which would be zeroised to represent value 0, the first of which would be

set to represent value 1, the first two of which would be set to represent value 2 and so on through to value n, when all attributes would be set. This approach enables differences in magnitude to be accounted for, although as with qualitative multi-state characters the degree of similarity or dissimilarity may be exaggerated, depending on the number of two-state attributes required to represent all possible values. In this particular case however, the association between individuals is exaggerated when negative matches are ignored. Thus, in the above example, if two individuals have values 1 and 2 respectively for the character in question these will be considered less similar with respect to this character than two individuals with values 2 and 3 respectively, as the latter case gives rise to two matches and the former to one. The choice of association measure is therefore critical in this case.

Using the non-additive approach each two-state attribute represents a different value and attributes are mutually exclusive, as in the above example given for continuous measurements. In the non-additive coding method described by Sneath and Sokal the first of each set of two-state attributes employed to represent characters is used to denote the presence or absence of the given character. This means that if two individuals possess a value for this character they will at least agree with respect to this first attribute, even if their respective values differ.

Although the additive and non-additive coding methods described by the above authors usually refer to ordered, non-metrical quantities, they are equally applicable to metric quantities and the additive coding method in particular has been useful in cases where binary representations are required.

(iv) Missing values

One of the problems in obtaining suitable numerical codes for character comparison is deciding on the treatment of negative score agreements between two individuals in finite data sets i.e. whether the absence of a particular characteristic in two individuals should be considered as contributing to the similarity between them. From the previous discussion it is obvious that this question becomes even more of a problem when characters are represented by sets of two-state attributes. Here the breaking-down process not only introduces redundant definitions which distort the degree of similarity or dissimilarity, it also has a weighting effect on the characters concerned, as the number of two-state attributes required for each multi-state character varies with the range of values the character may possess. As the exact effect this has depends on whether or not the mutual absence of characteristics in individuals is ignored, the choice of association measure in these cases is of vital importance. The different types of association measure which are capable of handling such two-state attribute sets are discussed below.

1.5.3 Estimation of Resemblance

Traditionally the measures of association used to estimate quantitatively the degree of resemblance between individuals coded in numerical form are referred to as coefficients of similarity or coefficients of resemblance. Sneath and Sokal<sup>46</sup> have suggested recently that the former term be restricted to coefficients which are strictly similarity coefficients to avoid confusion. As the interest in numerical procedures for classification increases several investigators have attempted to categorise

the different coefficients available and assess their suitability in different applications, but this has been difficult because of the wide range of measures available and the diverse nature of applications. Four basic types of coefficient are now recognised.

1.5.3.(a) Association coefficients

Some of the earliest coefficients of resemblance used in numerical classification are those which operate on a binary representation of the data. In the older literature these are usually referred to as coefficients of association. They are also often referred to as matching coefficients as they are based on counting the number of actual agreements between pairs of individuals compared with the number of possible agreements. These long established measures were used in a variety of disciplines before they were first adopted for use in numerical classification.

Many different association coefficients have been proposed and these large numbers have arisen mainly due to uncertainty over the treatment of negative attribute scores in pairs of individuals, and whether agreeing and disagreeing pairs of values should be treated equally. Thus, some association coefficients ignore negative attribute agreements altogether e.g. Jacchard's coefficient and Dice's coefficient, some give extra weight to matched pairs of values e.g. Dice's coefficient and others give extra weight to unmatched pairs of values e.g. the coefficient of Rogers and Tanimoto. One of the simplest association coefficients is the so called simple matching coefficient which gives equal weight to matched and unmatched attribute pairs and includes negative matches.

The different types of association coefficient may give widely different coefficient values for the same set of data. This in

itself is not a particularly serious defect but another characteristic of these coefficients is that they are not necessarily jointly monotonic i.e. the pair-wise associations between individuals ranked in increasing or decreasing order of magnitude do not necessarily lead to the same order of pairs in all cases. This could have a much more serious effect on the classification result. These differences have prompted a number of comparative studies of association coefficients attempting to define the relationships between them.<sup>86-88</sup>

Investigations so far have shown many of the association coefficients to perform closely e.g. the simple matching coefficient and the association coefficient of Rogers and Tanimoto have been shown to be jointly monotonic. Other association coefficients have also been found to behave similarly, although most of the investigations to date have concentrated on a small number of measures and difficulties of application have so far prevented a rigorous comparison of methods.

Coefficients of this type, which distinguish only between attribute values which match and those which do not are obviously most suited to qualitative data which can be meaningfully represented in binary form. They are therefore most appropriate for handling characters, which define a characteristic which is either present in an individual or is absent. The measures can however be applied to multi-state characters provided these are first suitably broken down into binary representations as described previously. This approach is not particularly suitable for multi-state characters which are either ordered or metrical, and especially if the data is continuous, because of the information loss in the transformation to binary form. Data of this type may be more realistically dealt with by similarity measures

capable of dealing directly with quantitative values. The approach is more suitable for multi-state characters which are purely qualitative, but even here difficulties may arise. One of the the major problems which has been discussed in a previous section is the weighting effect introduced when the ranges of possible values for each character differ. Also in these cases, leaving aside the problem of negative attribute matches, should matches arising in different characters be treated equally or should allowance be made for the fact that a match arising in say a six-state character is less likely than one which arises in a two-state character? Very few association coefficients take this into account and one of the better known matching coefficients which does consider the probability of a given match arising is the similarity coefficient proposed by Smirnov.<sup>78</sup>

A general association coefficient which is suitable for all types of data has recently been proposed by Gower.<sup>89</sup> One important advantage of this coefficient is that it is able to consider a mixture of data types in a single investigation.

#### 1.5.3.(b) Distance Coefficients

Distance measures are another group which have been extensively applied in numerical classification. They are most suitable for use with quantitative data and are often referred to in numerical classification studies as measures of 'taxonomic distance'. The different distance coefficients proposed, ranging from the earliest reported applications by Heinke to the present day have been well documented in the literature,<sup>40, 46, 74, 90-94</sup> and some of the important early formulations applied in the biological sciences, such as Pearson's coefficient of racial

likeness and the related generalised distance measure proposed by Mahalanobis have been mentioned in an earlier section. In general, distance measures express the similarity or dissimilarity between individuals in terms of their distance apart in an n-dimensional space whose coordinates are based on the characters used for the description of individuals. Most measures of similarity and dissimilarity, including distance measures give rise to coefficient values between pairs of individuals which satisfy the following conditions,

for individuals x and y

$$d(x,y) = 0 \text{ if } x = y ;$$

$$d(x,y) = d(y,x)$$

In addition to these, distance measures also satisfy the following condition, usually referred to as the metric or triangular inequality,

for individuals x, y and z

$$d(x,y) + d(y,z) \geq d(x,z)$$

This third property is an important one which distinguishes distance measures from many of the commonly used similarity coefficients. It also gives distance measures the property of being jointly monotonic.

The most commonly used distance measure which is now widely used in cluster analysis is the simple Euclidean distance, where the distance between two individuals i and j in an n-dimensional space is defined as follows

$$d_{ij} = \sqrt{\sum_{k=1}^n (X_{ik} - X_{jk})^2}$$

where  $X_{ik}$  is the value of the kth character for individual i.

This measure together with the problems associated with its use



have been fully discussed by Sokal,<sup>90</sup> Sneath and Sokal<sup>74</sup> and others.

Another metric closely related to the above is the absolute or city block metric,

$$d_{ij} = \sum_{k=1}^n |x_{ik} - x_{jk}|$$

This has been used by Carmichael and Sneath.<sup>95</sup> Other more general distance measures which include the Euclidean and city block distances as special cases are the Minkowski metrics.

One of the main difficulties in using distance measures of the above type is handling characters which are based on different unit measurements. These differing scales may seriously distort the overall degree of similarity between individuals, and to compensate for this and for differences in character ranges, standardisation procedures are often employed before distances are calculated. A variety of character standardisation procedures have been proposed.<sup>90,96,97</sup> The most common approach is to standardise characters so that they possess a zero mean and a unit variance, using the standard deviations derived from the complete set of individuals.<sup>90</sup> Although this has the effect of preserving relative distances, it and other similar methods have been criticised for also having the effect of diluting the differences between characters which may be of important discriminating value. Most investigators would agree that in the calculation of distance coefficients some form of preprocessing is usually necessary in order to calculate meaningful distances. However despite the widespread use of standardisation procedures at this time there is still very little general agreement over their validity and their exact effect on

the classification result.<sup>46, 74</sup>

The above distance coefficients are just a few examples of some of the more commonly used distance measures. Usually these coefficients can take any positive value and differ in this respect from other similarity and dissimilarity measures, which are usually normalised within the range 0 to 1, or -1 to +1. Another property of distance functions is that they may be easily transferred into a corresponding set of similarity functions, for example by considering reciprocal or complementary values which are normalised in some way. The reverse process of converting similarity measures to distance measures is more difficult, because of the need for distance measures to satisfy the metric inequality. However, a number of similarity coefficients can be put into metric form e.g. the simple matching coefficient, and various methods have been proposed for achieving this.<sup>49, 98-102.</sup>

In contrast with the examples given above some of the earlier proposed distance measures e.g. the Mahalanobis  $D^2$  statistic, have the advantage of allowing character correlations to be taken into account. However, some of the early measures have been designed especially for continuous data and many investigators feel they are unsuitable for clustering applications where discontinuous measurements are involved.

### 1.5.3.(c) Probabilistic Coefficients

The above measures of resemblance are the most widely applied in numerical classification. Recently much attention has been given to the problem of character value distributions and their effect on the classification result, and many probabilistic and information-theoretic similarity measures have been proposed which take these

into account. The coefficients which consider the frequency of occurrence of characteristics to be a measure of their importance, and the various problems arising in their use have already been discussed during the assessment of characters in section 1.5.2(b). So far these measures have not been widely applied because of the computational difficulties usually involved in deriving them. They are usually based on exact probabilities, although there are some exceptions to this, particularly in some of the earlier methods described e.g. Smirnov's coefficient.

In addition to the above probabilistic measures there are a number of measures which have their basis in information theory and consider the information content or entropy of characters. Information content in information theory is considered analogous to the concept of entropy developed in thermodynamics. Thus, it is considered to be a measure of disorder of characters and is directly related to the number of alternatives possible and their relative probabilities, when all known information is recorded. Shannon<sup>103</sup> developed some of the earlier ideas in this area and derived the following expression for characters which may take on a range of different values,

$$H = - \sum_{i=1}^n p_i \log p_i$$

where H is a measure of the uncertainty, or choice,  $p_i$  is the frequency attached to characteristic i and n is the total number of possibilities. Thus, assuming each character value for a given character is equally likely the entropy increases as the number of alternatives possible increases. If the probabilities of different character values are not equal then this has the

effect of lowering the entropy. It also has a weighting effect on characters, whose distributions are not equivalent, in much the same way as the probabilistic measures discussed previously. The recent literature shows that a number of investigators have considered this approach in deriving information indices for individual characters but so far studies have mainly been at an exploratory level and there have been few applications involving the setting up of similarity indices for complete individuals.

Information measures have a number of useful properties.

Because information statistics are based on probability theory, not only are they additive over characters they also allow the inter-dependency of characters to be taken into account, should this be required. A number of entropy measures have been proposed which take character correlations into account. In this case the probability distributions of characters are not assumed to be independent, and entropy values are calculated using conditional probability distributions. Several different approaches to this problem have recently been reported in the literature.<sup>104-107</sup>

Another advantage of the approach is that the relationship between the information statistic and the chi-square distribution enables statistical tests of significance to be applied. Finally, information measures like many association coefficients can be converted into metric form.

#### 1.5.3(d) Correlation Coefficients

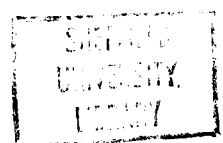
The last type of coefficient which has been widely applied in numerical classification is the correlation coefficient. These may be considered as special cases of a wider class known as angular coefficients.<sup>46</sup> The most commonly used coefficient of this type

is the Pearson product-moment correlation coefficient, which handles continuous data. Other similar coefficients have also been used for ranked and two-state characters. These measures have been considered over a wide area although many investigators have recently questioned their suitability for numerical classification.<sup>108,109</sup> For example, using Pearson's product-moment correlation coefficient the essential requirement for perfect correlation between two individuals is that the set of character values for one of the individuals be linearly related to the set of character states for the other, and the values of individual characters need not necessarily be in agreement. In some applications e.g. Wishart,<sup>110</sup> this coefficient has been found to be totally unsuitable for obtaining meaningful separations of the data, whilst in others e.g. Strauss et al,<sup>111</sup> it has been found more appropriate than other measures. Correlation coefficients are therefore still considered useful in some areas although in general they are becoming less popular than other measures.

#### 1.5.3.(e) Choice of Resemblance Measure

Many investigators have considered in detail the problems involved in choosing suitable methods of similarity and dissimilarity.<sup>40,46,74</sup> As seen in the previous section the choice of coefficient is often dependent on the nature of the original data. Thus association or matching coefficients are more suitable for use with qualitative information whilst distance measures are more appropriate for quantitative data and especially for continuous data. Where a genuine choice exists however there is at present no definite ruling as to which coefficient to use.

For example, in the case of dichotomous data the major problem is deciding whether or not to include negative matches in the similarity coefficient. In cases such as this the coefficient chosen is often the one which is shown by empirical evidence to be the most appropriate for the particular type of application in question. Another factor which influences the choice of method is the amount of computation involved. With improved processing facilities over the past few years this is becoming a less important consideration although with certain coefficients, where the computational load is high e.g. the probabilistic similarity measures, this is still an influential factor. The choice of resemblance measure is also influenced by the nature of the clustering method used. For example, in the case of the hierarchical clustering methods a number of investigators who have considered the mathematical properties of these have shown many of them to be unsuitable for use with coefficients which do not have strict numerical significance. Because of the arbitrary nature of scaling procedures and the processes involved in combining attributes, many coefficients have only ordinal significance and this means that quite a large proportion of resemblance measures are on mathematical grounds unsuitable for use with the majority of hierarchical clustering methods. For example Jardine and Sibson<sup>74, 112</sup> have shown that a very inaccurate representation of the data may result when using a hierarchical clustering method, such as the single-linkage method, unless a metric coefficient is used. However many disagree with their views and question the validity of a mathematical approach in choosing suitable methods. This point is discussed further in



in sections 1.5.4 and 1.5.5. Thus, with the general disagreement over the importance of the mathematical properties of methods, the resemblance measure chosen in cases where a choice of methods exists usually depends on evidence of previous performance, and on the particular preferences of the user.

#### 1.5.4 Methods of Cluster Analysis

As discussed in previous sections most classification methods based on cluster analysis proceed via a pair-wise measure of similarity or dissimilarity of the set of individuals in question. The wide variety of clustering methods available to handle these are well documented, but the different categorisations of methods appearing in the literature are often misleading. One of the problems of defining methods is that it is often difficult to draw clear divisions between the various classifications possible, and any one classification method may give rise to clusters which satisfy a number of separate conditions.

A good summary of the basic cluster arrangements possible is given by Jardine and Sibson.<sup>74</sup> These distinguish between simple and compound clusters, partitional and overlapping clusters, and numerically stratified clusters which are hierarchic or non-hierarchic. Clusters are defined as simple if no class includes any other class, otherwise they are compound. Partitional clusters, in contrast with overlapping clusters, are disjoint or else one cluster is totally included in another. Numerically stratified clusters are those which have associated numerical levels and clusters at a given level are nested with clusters at a higher level. If the clusters at each level are partitional then the arrangement is said to be hierarchical.

The various clustering methods currently in use in numerical classification applications are outlined below.

1.5.4(a) Clustering Methods and Applications

Partitioning techniques which give rise to simple clusters and form a partition of the set of individuals are often applied when the user is interested in partitioning the set into a predetermined number of clusters. These methods are also frequently referred to as optimisation techniques as they partition the set so as to optimise some predefined clustering criterion. One of the problems in using this approach is deciding on the appropriate number of clusters, and some of these methods allow this number to be changed during the course of the analysis. In addition they also allow re-allocation of individuals which may have been poorly classified initially.

The partitioning techniques which do not lead to distinct or disjoint clusters have become important in applications where it is essential to permit overlap between groups for the classification to be of any value. For example, in language studies overlapping clusters are particularly useful to account for multiplicity of word meanings, and in this area classification methods which give rise to overlapping clusters are usually referred to as clumping techniques. This term was first introduced by Jones and Needham and fellow workers at the Cambridge Language Research Unit<sup>113-118</sup> and the techniques they developed have been considered primarily for document retrieval purposes. Like most clustering methods clumping procedures begin with the computation of a resemblance matrix from the original data. The methods then seek to partition individuals into two groups so as



to minimise a cohesion function defined between the groups. The methods developed depend on the type of cohesion function defined and the smaller of the two classes obtained is usually the cluster sought. The clustering procedure is usually carried out by choosing a cluster centre at random and by successively eliminating individuals from this centre until the given cohesion function is minimised. In this way, by iterating from different starting points, a series of clusters may be obtained. However this independent search for classes not only gives rise to overlapping clusters but also to clusters which are not necessarily unique, and this is one of the most serious disadvantages of the approach.

Simple clusters and especially two-way classification schemes have become important recently for a variety of data analysis problems, including chemical applications directed towards property prediction.<sup>10,20,54,55</sup> Numerically stratified clusterings on the other hand convey more overall information about the data, and apart from the clumping procedures described above they are the methods most usually considered in applications where data structuring for retrieval purposes is of primary importance.<sup>2</sup>

One of the disadvantages of these methods is that they contain no provision for re-allocation of individuals which may have been poorly classified at an early stage in the analysis. This problem is more serious for monothetic than for polythetic techniques. Another problem is deciding on the number of clusters present. In many applications which use this approach this question is irrelevant but in cases where some indication of the number of groups present is required, a variety of statistical procedures have been

proposed for identifying the level in the hierarchy which gives the most meaningful separation of classes.<sup>119-122</sup>

Until recently numerically stratified clustering methods which give rise to disjoint clusters at each level have been more widely considered than those giving overlapping clusters. These are usually referred to as hierarchical arrangements and they are often represented graphically in the form of a dendrogram - a two-dimensional diagram illustrating the fusions or partitions which have been made at each successive level in the hierarchy. Initially hierarchical classifications were thought to be particularly relevant in biological systems, and the dendrogram representation closely resembles the traditional classification schemes developed in this area. Their widespread use in this area eventually led to their application in many different fields. However, it has been increasingly realised with the changing attitudes towards classification that nested mutually exclusive arrangements do not always provide the most satisfactory explanation of the data and within many of these areas there has been a move towards more flexible arrangements which involve overlapping hierarchies.<sup>74,123-125</sup> Despite this trend hierarchical groupings are still widely used and are thought to be the most practical, if not the most sensible arrangements in many applications. Computationally they are more straightforward and require less space than overlapping arrangements. This is an important consideration, for example in information systems, where large amounts of data are involved.

Hierarchical clustering techniques are usually implemented using agglomerative procedures, where individuals are successively grouped together until they belong to a single class. Divisive

procedures which operate in the reverse direction have also been used e.g. the techniques of association analysis developed by Lambert and Williams<sup>42,126</sup> and MacNaughton-Smith<sup>127</sup> but the agglomerative approach is preferred for a number of reasons. Firstly, it is more straightforward and easier to program. In addition, there is a greater risk with divisive techniques of inappropriate allocation of individuals. This is a particularly important factor in the case of hierarchical methods where, unless special reallocation procedures are employed, there is no facility for redistributing individuals once they have been classified. In the literature, hierarchical clustering methods are frequently categorised according to whether divisive or agglomerative procedures are employed, but Jardine and Sibson<sup>74</sup> have recently pointed out that such categorisations are invalid as they fail to distinguish between algorithms and the clustering methods which they implement. The basic differences between methods arise in fact because of the ways of defining similarity between groups, and between individuals and groups.

One of the first hierarchical clustering methods considered was the single linkage or nearest neighbour method.<sup>34,128</sup> This is computationally the simplest of the hierarchical methods and is one of the easiest to implement. It is so called because connections between groups and between individuals and groups are established by single links between pairs of individuals. Thus using an agglomerative approach groups, initially consisting of single individuals, are fused according to the 'distance' between their nearest members, at each level the groups with the smallest distance being fused. Distances between groups are therefore defined as the distance between their closest members.

Another similar approach where distances are based on furthest neighbours within groups is the complete-linkage method. One of the problems of the single-linkage method is the tendency of the method to cluster together at a relatively low level individuals linked by chains of intermediates. This so-called chaining effect is often viewed as a defect of the single-linkage method, as the majority of users are looking for homogeneous, compact clusters even though in general there is no reason to believe that these are the only types of structures present in their data. Jardine and Sibson<sup>74</sup> point out that this is more a description of the method than a defect and is useful when the user is looking for optimally connected clusters rather than homogeneous clusters. However, the effect does mean that the single-linkage method fails to resolve relatively distinct clusters if a small number of intermediate points are present,<sup>129</sup> and several modifications have been proposed in an attempt to overcome this problem. For example many of the density search techniques have their origin in the single-linkage method and arose in an attempt to overcome chaining.<sup>129-134</sup> These methods seek regions of high density or modes in the data, where each mode is taken to signify a different group.

The other important hierarchical clustering techniques which have arisen to avoid the extremes introduced by the single-linkage and complete-linkage methods are those which base the distances between groups, and between individuals and groups on some kind of average linkage procedure. A variety of different methods have been proposed for averaging over groups, and many of these were developed by Sokal and Michener.<sup>36</sup> These investigators also introduced weighting procedures to compensate for the possible

defects introduced by averaging over groups. In some applications these have been found to have an adverse effect on the classification result.<sup>135</sup> However, in other cases they could be useful to account for disparities in group size and sample size. Weighted and unweighted average linkage methods and their possible uses have been discussed in detail by Sneath and Sokal.<sup>46</sup>

#### 1.5.4.(b) Choice of Clustering Method

With such a wide range of clustering techniques available the choice of suitable methods is often a difficult one, in the absence of a formal definition of the term cluster. Occasionally, as in the case of resemblance measures, the choice of clustering method is governed by the amount of computation involved. For example, optimisation techniques usually require large amounts of computer time and are consequently unsuitable for use with large data sets. Hierarchical techniques have the advantage over these of requiring far less computing and are therefore more appropriate for use with larger data sets. However where there are no such clear indications of suitability, as for example when one of a number of possible approaches based on a given clustering technique must be chosen, this is a much more serious problem. In the past there have been very few attempts to resolve these difficulties. More recently with the growing number of possibilities, there have been attempts to evaluate quantitatively the performances of different approaches to make the choice of method an easier one. These have mainly taken the form of empirical investigations, but increasingly investigators are considering more formal approaches and some e.g. Jardine and Sibson,<sup>74</sup> and Wolfe<sup>136</sup> have attempted to construct a mathematical framework within which different clustering techniques may be investigated. Unfortunately the theoretical

considerations do not always support the empirical evidence and there is considerable disagreement over the feasibility of a theoretical approach. For example Jardine and Sibson<sup>74</sup> and Jardine, Jardine and Sibson<sup>137</sup> have objected to several hierarchical clustering techniques on mathematical grounds although in many applications these techniques appear to perform satisfactorily. Basically, these authors show that a clustering method which transforms a similarity or dissimilarity coefficient into a hierarchic dendrogram may be regarded as a process whereby the ultrametric inequality is imposed on the resemblance measure which may originally not have satisfied this particular condition. Then they specify certain simple conditions that any such transformation should satisfy and show that the single-linkage method is the only hierarchical clustering technique to meet all of these. In practice however this clustering method has been shown in certain applications to perform less satisfactorily than other hierarchical methods and this has led several investigators to criticise the technique and the mathematical arguments put forward in its favour. For example Williams et al<sup>138</sup> have found that the hierarchical techniques objected to by Jardine and Sibson are more helpful in providing useful information for the investigator than is the single-linkage method, and they question the need for these techniques to meet the proposed mathematical criteria. In another empirical investigation Forgey<sup>139</sup> concluded that the single-linkage method performed well with very distinct clusters of any shape, but as soon as a moderate number of intermediate points or 'noise' points were introduced the results quickly became erratic. Such conflicting viewpoints are commonplace in the current literature. Since classification techniques are used primarily for data simplification and data description many investigators, like the

above, feel that a pragmatic approach to the classification problem is a more reasonable one than one which restricts investigations to the use of mathematically acceptable methods. Others, however, feel that investigations into the theoretical aspects of clustering techniques are essential for a proper understanding of methods and that future investigation should be directed more towards the efficient implementation of existing techniques which are known to be useful and whose theoretical background is reasonably well understood, rather than towards the development of new approaches.<sup>140</sup>

#### 1.5.5 Evaluation Problems

At each stage in the classification process the user is faced with a wide range of alternatives, and with the rapid growth of methods in recent years there have been a number of attempts to put cluster analysis on a more formal basis and adopt a more rigorous approach to classification problems in general, as seen earlier. In addition to investigations into the theoretical basis of classifications for testing the validity of methods, on a less theoretical level many procedures have been proposed for testing the statistical significance of clusters.<sup>141-147</sup> However, whilst there continues to be no general agreement over the purposes of classifications and on what constitutes a cluster such procedures can only serve as a guideline in the evaluation of methods, and some investigators feel they may even be misleading. For example Williams and Dale<sup>40</sup> suggest that it is the usefulness to the investigator of the division into groups which is important, and that even if this division is not significant in the statistical sense it may still be of value to the user when large amounts of data are involved.

Traditionally classifications have been judged according to how well they perform in the particular application for which they have been constructed, and despite the move towards more formal evaluation procedures this qualitative approach continues to be widely applied. Indeed many investigators feel that because of the philosophical problems arising in defining classifications it will never be possible to develop precise mathematical models which are universally applicable, and that the ultimate criterion for evaluating methods will always be the value judgement of the user.<sup>148</sup>



## CHAPTER 2

The Structure Diagram in Chemical Information Handling and  
Structure - Property Correlation

Numerical methods have been used for some considerable time in chemical applications for the interpretation of experimental data. With the availability of more efficient structure handling techniques over recent years there has been a steady increase in the use of such methods, especially in areas concerned with identifying relationships in the data which could be useful for prediction or for establishing cause and effect relations. However, so far in such investigations explicit definitions of the structure diagram have not played a major role.

One area of application which is now receiving widespread attention is the investigation of quantitative structure-property relationships in biological systems, which could be useful for predicting the activity of compounds and helping to rationalise drug mechanisms. Empirical and semiempirical methods, which give approximate predictions of activity, have been most widely applied to date<sup>4-9</sup> and other more theoretical approaches based on molecular orbital calculations<sup>149-151</sup> which would be more valuable for establishing cause and effect relations have made slower progress because of the complex nature of processes taking place in biological systems<sup>152</sup> and the difficulties of calculating accurate wave functions. In the more empirical approaches developed, physicochemical characteristics of molecules have generally been considered more useful parameters for activity rationalisation than structural characteristics, and fewer investigations have been based on structural attributes derived directly from the structure diagram.

The empirical and semi-empirical methods developed for use in structure - property studies to date have been based largely on statistical correlation techniques which seek direct relationships between structures and properties. More recently a variety of non-parametric pattern recognition techniques have been applied, and the structural attributes of molecules have been more widely considered in this type of application.<sup>5,12,13,54,55,153,156</sup> These methods first received attention in spectroscopic studies<sup>157</sup> but they are now gaining in popularity in biological applications, because of their ability to handle diverse structural types, and property measurements of a non-parametric nature.

Methods of pattern recognition could be put to a variety of different uses in chemical structure applications although so far the main emphasis has been on the development of methods which could be useful for property prediction. Another important possibility is the development of more efficient file-handling strategies in large structure collections. Classifications which are suitable for this type of application have been proposed in other areas<sup>1-3</sup> and it is possible that these could be used in large chemical information systems to improve file-handling techniques and to reduce the cost of retrieving compounds for drug research.<sup>158</sup> As discussed below the structure diagram could be a valuable tool in this type of application.

## 2.2 The Structure Diagram

### 2.2.1 Use in Chemical Communication

The structure diagram is one of the most widely used representations of chemical compounds. A large proportion of the chemical literature is structure-oriented<sup>159</sup>, as are many of the questions posed by

chemists. As a result of this many chemical information systems are structure-based, and in automated systems, processing is usually carried out on a machine-readable representation of the structure-diagram.<sup>16</sup> This is often fragmented to ease handling problems and the various fragmentation methods which have been developed have proved to be very useful for search and retrieval as well as for storage.<sup>15,16,160</sup> If similar representations could be used in the development of automatic classification or quantitative structure-activity methods, these methods could be of very general applicability in chemical information systems.

### 2.2.2 Structure-Property Relationships

The structure diagram may be considered as a very approximate two-dimensional projection of the real structure. It can also be thought of as an approximate pictorial representation of the wave function of the molecule. Usually it corresponds more closely to valence bond or localised molecular orbital descriptions than to delocalized molecular orbital descriptions, because of the difficulty of representing delocalised bonds in graphical form. However, it can distinguish between  $\sigma$  and  $\pi$  bonds, and takes some account of electron delocalisations by equalising alternating single and double bonds.

In addition to providing an approximate method of indicating bond orders the structure diagram also shows the atoms which are connected, and although it does not provide explicit descriptions of atom arrangements and stereochemical relationships these are sometimes implicit in the two-dimensional definition. Because of these relationships and the relationship between the wave function, structure and other properties, the characteristics of structure

diagrams should be correlated approximately with the physical, chemical and biological properties of the molecule it represents. It is difficult to estimate the extent of this correlation 'a priori' because of the large approximations involved, but the expected imperfect correlation between structure diagram and observable properties could be of some value in structure-property studies, and could have other uses in automatic classification applications. One of the problems of using automatic classification methods, as seen in Chapter 1, is devising suitable procedures for measuring classification performance, and here the correlation could be useful for simulating the predictive performance of classifications and the similarity measures used to derive them. The extent of the agreement between observed and 'predicted' properties could then provide a useful basis for the comparison of methods. The expected correlation also leads to the possibility that resemblance measures or classifications based wholly or in part on structural features will be of some value for property prediction.

### 2.2.3 Use in Structure-Property Studies

From a very early stage in the development of empirical structure-property methods the structure diagram has been considered of limited value for correlating changes in chemical structure with changes in biological response. Even before the beginning of this century investigators were starting to use physical and physicochemical properties in preference to structural formulae.<sup>161-163</sup> In present day applications these parameters are still often preferred to structural data, and, with the exception of the Free-Wilson regression model<sup>7</sup> and some recent studies based on pattern recognition, most of the investigations in this area are based on the semi-empirical regression

model developed by Hansch<sup>4</sup>, which considers linear free energy-related parameters from organic chemistry.

The usual arguments raised against the structure diagram in structure-activity studies is that it is not sufficiently discriminating to resolve the important properties of chemical compounds which govern biological activity. A drug response is usually the outcome of a complex series of events which control the transference of the drug to the biological receptor and the reaction taking place at the receptor site, and most investigators feel that it is preferable in this type of application to consider the physical and chemical properties of structures which are most likely to influence these events.

The questionable suitability of the structure-diagram in structure-property studies has been discussed at length in the recent literature<sup>8,164-166</sup>. For example Cavallito<sup>164</sup> has stressed the need for steric and electronic descriptions of component functional groups and of the molecule as a whole, and points out that these features cannot be obtained satisfactorily from two-dimensional or even three-dimensional descriptions of the structure diagram. Seydel<sup>165</sup> also believes that physicochemical properties are more useful, and that an understanding of drug action cannot be obtained by simply comparing structural formulae. Other similar views are held by Albert<sup>166</sup> and Verloop<sup>8</sup>, and Albert supports his arguments by giving examples of drug activities where the necessary physical properties to bring about a biological response can be produced by many different kinds of structure. However, where there are no obvious links between structure and biological action a number of separate structural requirements may be involved and it is possible that a careful exami-

tion of individual substructures in this case could help identify the characteristics most important for activity.

Although physicochemical properties of structures have been preferred, the semi-empirical methods based on the Hansch approach have until recently met with limited success. The basic Hansch approach is modelled on the Hammett equation<sup>167</sup>, originally developed for use in nonbiological systems, and Hansch<sup>168</sup> has suggested that one of the reasons for the moderate success of the approach is the reluctance of users to explore parameters other than the well known Hammett  $\sigma$  constants. These were originally used to investigate the effects of substituents on reaction rates in organic chemistry, and although they are important for explaining activity at the biological receptor they do not account for other important processes taking place such as membrane penetration and transport to the receptor site. In the early sixties Hansch<sup>169</sup> introduced lipophilic parameters into the basic linear free energy-related model to account for some of these processes, and in more recent applications steric parameters and an increasing number of physical properties have also been explored which may influence the changes taking place.<sup>4,6,8,170,171</sup>

In addition to the above, structural descriptions have been introduced in a few cases. For example, Hansch and Yoshimoto<sup>168</sup> have used structural information with linear free energy related parameters in an investigation of a series of benzamidine derivatives, and have shown these to improve the agreement obtained in this case between structure and property. In another application Martin et al<sup>172</sup> have combined structural and physical properties to investigate the relationship between the structures of a group of aminotetralins and aminoindanes and their inhibitory

properties, using discriminant analysis. This very recent interest in structural parameters has possibly been influenced by the recent success of structural descriptions in pattern recognition applications, and the investigations in this area has served to narrow the gap between existing empirical and semi-empirical approaches.

#### 2.2.4 Comparison of Structural and Physicochemical Parameters in Structure-Property Studies

Despite the approximations in the structure diagram, structural features in empirical structure-property studies have a number of possible advantages over semi-empirical correlation methods based on experimentally determined physicochemical quantities. One factor which limits the value of a physicochemical approach is the reliability of the experimental quantities involved. For example, the physical constants used in the linear free energy related model are usually obtained from non-biological systems, and one of the problems here is ensuring that the experimental conditions are satisfactory. In addition, because of the difficulties of obtaining accurate measurements the parameters required are often in limited supply. Partition coefficients, which have been widely used since the importance of hydrophobic bonding properties was first realised, are an example of this. The difficulties of obtaining reliable values of this parameter has been partly overcome by the discovery that partition coefficients have additive and constitutive properties which enables them to be calculated from the individual contributions of the molecular components.<sup>173-177</sup> The value of this however is seriously limited by the accuracy of the available experimental data, and the fact that experimental values are often determined under different conditions. Also experimental values used for the calculation of



new partition coefficients are often obtained from different systems, and because they do not account for steric and electronic interactions with neighbouring groups, the parameters derived from one system may not be suitable for use in another.<sup>4,178</sup> For example it has been shown that partition coefficients determined in aromatic systems are not suitable for use in aliphatic systems,<sup>173,175,176</sup> and the deviations which have been observed in this case are thought to be due to folding interactions in aromatic systems, causing depressions in coefficient values. Interactions with neighbouring groups are now being studied more closely, but until more is understood about these, measured coefficients are preferable to calculated ones, particularly in larger, more complex structures. Another difficulty with partition coefficients is that they are usually parabolically and not linearly related to the biological response and, although useful approximations can be made by assuming a linear relationship, this has discouraged a number of workers from using partition coefficients as reference systems.

These various considerations therefore limit the applicability of the method, whereas similar limitations do not arise with structural parameters, and these have the advantage of being able to handle larger molecules for which experimental parameters are not yet available. The only restriction in this case is the availability of reliable biological properties for a sufficient number of structures.

One of the main arguments put forward in favour of the semi-empirical approach is that the parameters used have a physical meaning and it is possible to give the regression equations a physical interpretation which might help towards an understanding of

action. However, the value of the approach as a diagnostic tool depends on the relationships arising between parameters, and, as larger numbers of experimental constants are introduced into the basic linear free energy-related model, the interpretation of the regression results becomes increasingly difficult. Usually the inclusion of larger numbers of variables has the effect of improving the overall agreement obtained between structure and activity, and this is important from a prediction point of view. However, many of the semi-empirical constants now used have been found to be interdependent, which makes it difficult to interpret individual contributions when they are applied simultaneously. Another problem arising is that many of the physicochemical constants have themselves been found to be made up of several components,<sup>179,180</sup> each of which may have a different influence on the structures reactivity. Because of these difficulties it has been suggested that the physicochemical model may be of limited value in establishing casual relationships between structures and properties, and that such mechanistic studies should be left to more specific, quantum-mechanical parameters.<sup>8</sup>

Structural parameters are considerably more straightforward, and, although a physical interpretation of the result is not essential in this case, it is possible that the individual contributions of structural components to activity will be of some help in rationalising drug action.

#### 2.2.5 Choosing Suitable Substructures

One of the main problems in using structural parameters is deciding which features of the molecule are the most important. The seriousness of this problem depends on the particular approach used. Thus, in the Free-Wilson mathematical regression model biological activity

is related only to the substituents that vary within the series. Relatively simple substituents are usually considered, and these are not broken down into smaller components. Other structure-property investigations based on the structure diagram have taken into consideration the whole molecule, but these have usually concentrated on the structural features considered to be of greatest chemical significance. This applies to the majority of supervised and unsupervised learning methods developed so far. The approach is not an entirely satisfactory one, however, as it prejudices the value of substructures, and risks overlooking features of possible interest. A systematic analysis of the structure diagram reduces this danger, but in this case it is difficult to extract substructures which reveal all the groups of chemical interest. For example, if only large substructures are chosen they may mask important functional groups. Smaller substructures on the other hand often miss important information on ring systems such as, for example, the relationships between substituents, which may be an important factor determining a structure's behaviour. These difficulties could be overcome if all possible substructures were included in the description, but this would be impractical because of the amount of redundant information involved.

Representations with some redundancy are acceptable in certain types of application, for example in substructure search systems where screen strategies operate by matching query and structure representations to establish whether particular substructures present in the query are also present in the structure. However, if retrieval is based on measuring the degree of association between  $query^r$  and structure representation, or if structure comparisons are required

for classification or prediction purposes, then the redundant information could seriously distort the levels of similarity and dissimilarity obtained. The problems of holding redundant information when applying similarity coefficients have already been discussed in detail in Chapter 1. Unfortunately, it is not certain exactly what effects redundancy has on the performance of methods. In this particular case it is possible that the additional information will be of some value in the classifications, although excessive redundancy is expected to have an undesirable effect. These questions have been largely avoided in chemical applications so far.

Similarly, in regression analysis redundant substructural descriptions could cause difficulties. Firstly too many parameters may be involved and secondly many of the substructures may be too highly correlated to be included in the same regression equations. The problems arising with interdependent fragments in regression analysis are discussed in more detail in section 2.5.

Another difficulty arising is that the importance of substructures may depend on the particular objectives of the study, and whether for example the analysis is required for the prediction of a particular biological property or to organise chemical structure data for retrieval purposes. The problem is therefore to extract from the structure diagram the structural features which are most relevant to the analysis in question. In regression analysis there is the additional problem of ensuring the number of substructures employed does not exceed the number of compounds, and it has been suggested that some of the recently developed methods of

pattern recognition could be of some value in this area.<sup>5,12,55,181</sup>  
However, the supervised learning methods proposed to extract the substructures most relevant to the analysis have been questioned in the recent literature.<sup>182</sup>

### 2.3 Automatic Classification Methods

As mentioned in earlier sections the main interest in automatic classification methods in chemical applications to date has been to identify relationships between structure and property data which could ultimately be of value for property prediction. In many areas, and particularly in biological activity studies the available data is of a qualitative or semiquantitative nature, measured either on a nominal scale or on ordinal scale, indicating relative degrees of activity. The question most usually asked therefore is whether or not a given structure is likely to have a particular activity or level of activity, and the problem is one of developing classification rules which can successfully separate structures into one of a number of fixed, predetermined classes. As a result most of the classification methods developed have been based on the supervised learning approach. The main interest so far has been in developing methods which can discriminate between active and inactive compounds and a variety of techniques have been developed specially for handling this type of two-class problem. These are outlined below and other less widely applied methods, such<sup>as</sup>/visual display techniques and unsupervised learning methods are also summarised.

#### 2.3.1 Supervised Learning

The basic aim of supervised learning methods is to develop classification rules which can correctly classify the data for which properties are available and to subsequently apply these rules to

categorise individuals for which the required property is not known. In this process the initial data is referred to as the training set, and the data undergoing classification as the test set.

A method which has been widely applied in chemical structure applications for the analysis of spectroscopic data is the linear learning machine<sup>157</sup> and this has more recently been considered in a number of structure-biological activity studies.<sup>12,55,153,183</sup> Using this approach error-correcting procedures are employed to define a linear function which can successfully separate the structures in the training set into active and inactive categories. The defined hyperplane, often referred to as the 'decision surface' is then used to predict the likely activities of the test set structures. As in most non-parametric pattern recognition applications reported in the literature to date the classification rules operate on a matrix of coefficients defining relationships between pairs of structures, and in the case of supervised learning applications these relationships are usually defined in terms of a distance function. Preprocessing procedures are then often applied to this distance matrix to transform the originally defined attribute space into a form more suitable for classification. For example, data transformations are often employed to make class discrimination easier. They are also employed to reduce the dimensionality of the original n-space to ensure that the ratio of the number of structures to the number of structural features is within acceptable limits.<sup>72,73</sup> These reductions are achieved either by discarding dimensions, considered expendible, or by combining them.

One of the main limitations of supervised learning methods of this nature is that the methods are only as good as the initial training set. The preprocessing stage necessarily imposes some bias on the method, and if the training set does not adequately represent the structures to be classified there is a serious risk of misidentification. Also, using the learning machine it is often difficult to separate the preprocessing stage from the decision-making stage. In the case of feature selection, for example, the importance of features is usually determined by introducing each dimension individually or in a group and measuring their separate effects on the classification result. Another difficulty is that preprocessing requirements often oppose each other and in this case a compromise must be reached.

Discriminant analysis is another supervised learning technique which has been used for classification in chemical structure applications.<sup>12,20,153,172</sup> This approach is similar to the linear learning machine in that it seeks a discriminant function which can be used to place structures into one of two categories. The same theoretical limits on the structure to feature ratio are also applicable. In this case, however, the discriminant function is trained by the method of least squares and these procedures have a firmer statistical basis than the feedback learning procedures employed in the learning machine case. This gives the approach a number of statistical advantages but less flexibility.

An alternative to the learning machine which has been considered by a number of workers is the K-nearest-neighbour classification method.<sup>20,54,55,153</sup> Individual relationships between structures are computed in the usual way, and the approach assumes that the closer two points are in the defined structure space the more alike they are

from a property point of view. Structures of unknown activity are therefore classified by the majority rule of the K nearest known structures in the training set. The approach is preferred by some investigators because of its conceptual and computational simplicity and the fact that many of the measurements involved have a firm statistical basis. It also gives an indication of the overall relationships arising between structures, which the learning machine does not do, and there are no restrictions on the number of structural features which can be used in the analysis.

### 2.3.2 Visual Display

The objectives of ordination or display methods is to reduce the structure space to a small number of dimensions, so that the ultimate classification can be performed by the user. The different approaches have been outlined in Chapter 1.

In chemical structure applications linear and non-linear mapping techniques have received the most attention, and in a recent investigation of classification methods in chemical applications, Kowalski & Bender<sup>20</sup> discuss the relative merits of these two approaches. The basic difference between them is that the final coordinates in the non-linear case are not linear combinations of the original n-space. These workers suggest that the non-linear approach is a more useful method of dealing with multivariate chemical data and that the non-linear, error minimisation mapping procedure proposed by Sammon<sup>184</sup> is possibly the most useful approach, because it attempts to preserve interpoint distances.

The validity of the final projection depends on a number of factors, including the type of resemblance coefficient used, and ideally these should be monotonic. Also, because the projections are only



approximations of the original n-space, and interclass boundaries are not exactly defined the above investigators suggest that structures appearing near the interface of two classes should be classified by other means.

Display methods have not received much attention as a classification method in their own right, and their main use has been in providing an aid to other more accurate methods, by allowing the user a rough visual examination of his data.

### 2.3.3 Unsupervised Learning

Methods of finding clusters in multidimensional data using unsupervised classification techniques have been considered for some time, although very few applications of this approach have been reported in the chemical literature. Some recent applications have appeared for correlating structure and property data. For example Kowalski and Bender<sup>20</sup> have applied a hierarchical clustering technique to identify clusters in two synthetic data bases of chemical interest, and they compare the usefulness of this approach with display and supervised learning methods. Sneath<sup>10</sup> has also used a hierarchical cluster method to classify a group of naturally occurring amino acids, using structural descriptors based on the structure diagram, and he uses the relationships obtained to correlate the structures and biological activities of a group of peptides.

The interest in unsupervised classification methods is growing steadily, and, as in the supervised case investigators are beginning to consider their value in preliminary investigations for extracting the most relevant material for more accurate studies. Hansch,<sup>181</sup> for example, has used a hierarchical clustering technique in conjunction with regression analysis to cluster the substituents of a closely

related series, so that the most suitable derivatives could be selected for regression analysis. Other possible uses of this approach are discussed below.

#### 2.3.4 Comparison of Supervised and Unsupervised Learning Approaches

Although very few classification methods have been applied to chemical structure data the techniques already used in this area show these methods could be of some value in structure-property studies, in cases where more accurate statistical correlation techniques are inapplicable. So far supervised learning methods have been developed which have been very useful for handling dichotomous variables. The less widely used unsupervised learning approaches have the advantage that they can also handle more accurate property measurements, since the property undergoing prediction in this case is not directly involved in the development of the classification rules. They could therefore be of use in areas where supervised approaches are inappropriate and the data is not sufficiently accurate or the conditions are not suitable for regression analysis.

Both supervised and unsupervised approaches have a number of useful properties. In the unsupervised case there are no theoretical limits on the number of structural attributes which can be included in the analysis. There are also fewer preprocessing requirements compared with the supervised approach, although some preprocessing may be needed in this case, for example, scaling procedures to prevent inadvertent feature weighting when measurements of different units are employed. One of the possible benefits of the preprocessing procedures applied in the supervised case to obtain a better separation of the data is that these are often thought to provide useful information

on the relative importance of individual substructures. Unsupervised methods on the other hand give a better indication of the overall relationships between structures.

Another important property of unsupervised methods is their possible application in large data collections for storage and retrieval purposes. Investigations in this area on a number of document based collections have shown that stratified and hierarchical cluster methods can be applied on a small scale to develop retrieval strategies which are more efficient than linear retrieval methods and also potentially more effective in terms of precision and recall.<sup>2</sup> If suitable large scale procedures could be developed in this area then it is possible that such file arrangements could lead to considerable retrieval benefits in chemical information system currently operating in the registration and substructure search modes.<sup>15</sup>

#### 2.3.5 A Novel Classification Method for Handling Chemical Structures

In view of the possible uses of unsupervised classification methods in chemical information systems, a method has been developed for the classification of chemical structures which combines a hierarchical clustering method with some automatic structure handling techniques of very wide applicability in existing computer-based systems. The main objective has been to develop methods for handling the structural features of chemical species, where these are derived automatically from a connection table representation of the structure diagram.

This is one of the first investigations reported which bases the classification of structures on a systematic analysis of the structure diagram, and the possible advantages of using this approach have been discussed in section 2.2.5. It is also the first unsupervised

classification method reported which attempts a more formal approach to property prediction, using a technique similar to the K-nearest neighbour method.

The classification methods developed are described in detail in the following chapter.

## 2.4 Regression Analysis

The empirical and semi-empirical regression models developed so far to investigate structure-biological activity relationships aim to optimise activity within groups of related structural types by considering variations in side-chain structures. The structures under investigation are expected to share a common nucleus, which is assumed to have a constant effect on the result, and changes in the observed biological response are attributed to the physicochemical or structural properties of the variable part of the structure.

The two main approaches developed are the linear free energy-related model developed by Hansch and the mathematical model, usually associated with Free and Wilson, both of which have been referred to in earlier discussions.

### 2.4.1 The Semi-empirical Model

The semi-empirical parameters used in the linear free energy-related method have been discussed in section 2.2.4. Using this approach, biological activity is correlated with one or more physicochemical properties with which the particular drug response is assumed to be associated. The very early attempts to define semi-empirical relationships between structures and biological activities, using Hammett reaction rate constants were largely unsuccessful. Eventually it was realised that a biological response is not necessarily governed by chemical reaction rates and other processes, such as drug transport and membrane penetration, may be of dominant importance. To help

account for these processes Hansch and co-workers introduced a substituent partitioning parameter into the basic Hammett expression in 1964.<sup>169</sup> This parameter was defined as the difference between the logarithms of the octanol/water partition coefficients of the substituted and unsubstituted parent compound in a series. The expression they derived led to a considerable improvement in the agreement between structure and activity, and it is now often referred to as the ' $\rho$ - $\sigma$ - $\Pi$ ' equation'. Since its initial formulation this equation has frequently been modified by introducing steric parameters and, more recently, a range of experimental parameters (see earlier) to help improve the agreement obtained. The use of the Hansch method and modified versions of it have rapidly increased within the last ten years, although as discussed in previous sections users have been slow in considering parameters other than those involved in the basic Hansch model.

#### 2.4.2 The Empirical Model

The empirical, mathematical model, in which biological activity is expressed as a function of the activity contributions associated with substituent groups and the parent compound, has received much less attention than the Hansch approach.

Serious work on empirical, mathematical methods began as early as 1956, when Bruice and co-workers<sup>185</sup> constructed a mathematical model to correlate the thyroxine-like activity of a group of congeners with the sum of constants assigned to different substituents of the molecules. They obtained reasonably good correlations between observed and calculated activities. Free and Wilson gave a more general description of this empirical model in 1964,<sup>7</sup> where they defined the biological response of a derivative in a homologous series in terms

of the sum of the substituent group contributions to activity plus that of the parent structure. The substituent contributions are evaluated in much the same way as the physicochemical parameters in the semi-empirical model, i.e. by the least-squares solution of a set of linear equations, one for each of the molecules in the series. The basic assumption in this case is that every time a particular functional group appears at the same position in the molecule it will add or subtract a constant amount to the overall biological activity of the molecule, regardless of what other substituents are present. The interpretation of substituent constants and the ability of these to predict the activity of any combination of substituents will therefore depend on the validity of this additivity assumption. The method has been applied by a number of different investigators,<sup>186-188</sup> although so far applications of this approach for the design of new lead structures have not been reported.

The success of these two regression methods depends on the members of the group under study having a similar mode of action, and to increase the likelihood of this it is important to keep structural differences to a minimum. The most serious limitation of the mathematical model is the need for activity contributions of substituents to be additive. The practical limitations in the physicochemical model have already been discussed in section 2.2.4. These various requirements limit the predictive value of the methods and the need to minimise structural variations also limits their value as diagnostic tools for rationalising activity.

2.4.3 A Novel Empirical Regression Model Based on Explicit Structure

Definitions

So far, regression methods have not been used to relate property or activity data to total molecular structure. If suitable methods could be developed in this area this would remove the restrictions on structural variations arising in existing approaches, and the analysis could be used to explore a much wider range of structural types. This, in turn, would increase the use of the method as a tool for prediction and it may also increase the value of the regression solutions for interpreting the role played by individual substructures.

In view of these possibilities a method has been developed for correlating biological activities and other properties to the characteristics of the entire molecule. The method is empirical but differs from the Free-Wilson approach and the other regression methods developed to date in making no distinctions between side chains and parent structures, and breaking these down in an equivalent manner. Structure diagrams are fragmented systematically using some techniques of chemical structure handling important in existing storage and retrieval systems, and like the classification approach described above, the method developed could therefore be useful in computer-based systems where properties and structure diagrams are already available in machine-readable form.<sup>22-24</sup>

Perhaps the most valuable property of this approach compared with the Hansch and Free-Wilson regression models, is its ability to handle diverse structural types, as this means that the method could be used to explore possible new lead structures.<sup>5</sup> So far in structure-

biological activity studies the main approach to discovering new classes of biologically active compounds has been to investigate the chemical changes taking place at the molecular level, and in particular the changes occurring at the biological receptor. As discussed in the introduction to this chapter promising new approaches are now being developed in this area, but investigations have been difficult because of the complexity of drug-receptor interactions and the largely unknown nature of biological receptors. An empirical regression method able to examine total structure could be useful in this area for speculating on new leads, and possibly for initiating more direct studies.

Details of the method developed in this study are given in Chapter 4.

## 2.5 A Comparison of Regression Analysis and Pattern Recognition in Quantitative Structure-Property Studies

In structure-property investigations to date, the choice of methods has usually depended on the type of property measurements available. Thus, parametric approaches such as regression analysis have been considered in cases where suitable interval or ratio data is available, whereas non-parametric classification methods have been used when only quantitative measurements are available. As more attention is given to non-parametric methods, and classification methods are introduced which are capable of handling more accurate property measurements the user is faced with a growing number of alternatives,



and it is necessary for him to consider more closely the relative advantages of each approach.

One of the main advantages of the classification approach is that there are fewer requirements concerning the underlying statistical nature of the data, and the only assumption needed in this case is that a relationship between the structures and the defined properties exists. Another advantage is that the interpretation of the data is not restricted to current accepted schools of thought. In contrast, the regression method assumes certain relationships to exist 'a priori'. This places constraints on the method and on the interpretation of the data, but the theoretical basis of the approach also gives it a number of advantages over the classification approach.

#### 2.5.1 Data Requirements

##### 2.5.1(a) The Dependent Variable

One important requirement of the regression method is the need for property measurements, considered as the dependent variable, to be quantitative, and ideally these values should be measured on a continuous scale. Semi-quantitative measurements are also suitable, provided degrees of activity are measured on a suitable interval scale, and the range of activity covered is large enough. However, measurements of this type are not often available, because of the practical difficulties involved in measuring accurate response rates in biological systems, and this is seen as one of the major limitations of the approach at the present time. Classification methods on the other hand can handle less accurate property measurements and can deal with the qualitative and semi-qualitative measurements more usually found in this area.

2.5.1(b) Independent Variables

Another restriction of the regression method is the need to control the number of explanatory or independent variables, used in the analysis, and to keep this number below the number of structures involved. This could be a particularly serious problem in cases where diverse structural types are considered or when investigations are based on a systematic analysis of the structure diagram, because of the larger numbers of substructures usually involved in these cases. Similar restrictions arise in the supervised learning case but they are not applicable in unsupervised methods.

The other important factors influencing the agreement obtained between structure and property data in the regression case are the number of degrees of freedom involved and the correlations arising between structural components. Although fewer restrictions of this nature are expected in the classification case, the advantage of the regression approach is that there are available reliable statistical criteria to measure the exact effects of these various conditions on the agreement obtained. In the classification case the uncertain mathematical properties of the approach have so far prevented a rigorous assessment of the conditions which may influence the final result and those which do not. Thus in the basic regression model the number of degrees of freedom have a measurable effect on the result and character correlations are known to increase the standard errors of the regression coefficients. The problem with character correlations in the classification case is further complicated by the fact that certain types of correlation are thought to be essential for a meaningful result, and considerable difficulties arise over deciding which

correlations are admissible and which are likely to have an adverse effect on the final result. These problems have been discussed in detail in Chapter 1.

Choosing a suitable numerical representation of the structural attributes is also more of a problem in classification. Quantitative or semi-quantitative representations are most suitable in the regression case but in the classification case, depending on the way substructures are initially defined and the type of similarity coefficient applied, a variety of qualitative and quantitative representations are possible which may be equally appropriate.

## 2.5.2 Evaluation Procedures and Available Statistical Criteria

### 2.5.2(a) Regressions

The available statistical procedures in regression analysis make it possible in this case to test the reliability of the results and to evaluate the contributions of individual substructures. This is the most important advantage of the approach but the validity of the available statistical criteria depends on the data first satisfying a number of important distribution requirements, in addition to the data requirements already outlined above.<sup>189</sup> Thus, the basic regression model requires that the explanatory variables be measured without error and that the property parameter considered as the dependent variable be taken from a population of independently and normally distributed variates. The first of these requirements is no problem in structure-property studies based on two-dimensional substructural descriptors, but the second condition is much more difficult to satisfy. Practical difficulties usually prevent biological experiments being repeated a large number of times. The biological

parameters used in regression analysis are therefore usually the outcome of only a few independent observations and it must be assumed that these values are based on populations satisfying the above data requirements. It is also assumed that the variance around the regression line is constant and independent of the explanatory and dependent variables.

Provided these various conditions are approximately met another problem arising is the misuse of the available statistical quantities.<sup>8,190</sup> There are a variety of statistics available to estimate the significance of the correlation and the correct usage of these often depends on the particular interests of the user. For example, the standard deviation of the estimate is a useful indication of whether the relationship provides a good summary of the data, but it does not take into account the numerical range of the dependent variable and is not a reliable indication of the extent of the agreement obtained. The multiple correlation coefficient ( $R$ ) is more useful for this purpose and the literature shows this to be the most frequently consulted statistic for estimating the significance of the correlation. However, as a 'goodness of fit' measure it is not the most reliable statistic because it does not take into consideration the number of degrees of freedom in the analysis. High correlation coefficients obtained using large numbers of structural attributes therefore do not necessarily mean that the agreement between structure and property is significant, because the successive introduction of explanatory variables will tend to increase the value of  $R$ . Because of this,  $R$  is also inappropriate for comparing the performance of two regressions which involve different degrees of

freedom although it has been used for this purpose. To estimate the significance of correlations,  $R$  should therefore be considered in relation to the number of structures and the number of independent variables, and significance levels should be estimated from these values using the  $F$  test. Statistical measures are also available to compare regression coefficients and to test whether the values obtained are significantly different from each other. However, the correlations reported have often omitted significance tests, and very little attention has been paid to relationships arising between independent variables.

#### 2.5.2(b) Classifications

Similar statistical guides are not available in the classification case, although occasionally null hypotheses have been advanced which enable certain statistical tests to be applied to estimate the significance of clusters.<sup>191</sup> Usually it is assumed that the individuals studied belong to a single class or that they are regularly or randomly distributed with no class identity. The difficulty in using this type of approach is that the usual 'goodness of fit' tests such as chi-square and the Kolmogorov-Smirnov statistics may not always be suitable. They can be applied to supervised learning problems, as here a certain result is expected, and they can therefore be used to estimate the success rate of the method. However, they are of questionable value in unsupervised classification applications and in this case it is necessary to find simpler measures to estimate method performance, such as for example criterion functions. These have been discussed in Chapter 1.

Quantitative methods of evaluation are still the exception rather than the rule in classification applications and, as stated in the previous chapter, classifications are still largely evaluated empirically on the basis of their performance in the particular application in question. In structure-property investigations therefore performance is usually measured in terms of predictive power.

### 2.5.3 Property Prediction

So far there have been very few structure-property applications reported in the literature which have been used to predict the biological activity of a compound before its synthesis. This is possibly because of the potential economic value of such predictions. However, until more recently there have also been very few simulated predictions reported and investigators have relied on less accurate criteria to estimate the predictive utility of methods.

In supervised learning applications the suitability of methods for prediction has been judged largely on their ability to correctly classify the initial training set structures. Similarly in regression analysis very few of the correlation studies carried out to date have reported on the predictive value of methods, and most investigators have taken the correlation obtained from the analysis to be a sufficiently useful guide to the predictive power of the regression equations. However, in each case the property values estimated from the initial analyses are not necessarily a good indication of predictive value as each structure included in the analysis is allowed to influence its own result. A more reliable estimate of predictive power can be obtained using 'hold n out' procedures<sup>5,19</sup> and these

have been applied in some more recent investigations.<sup>12,20,55,172</sup>  
In using this technique a compound or set of compounds is excluded from the original analysis and the regression solutions or classification rules developed are then used to predict the activity of the structure or structures excluded. The procedure is usually repeated until all the structures have been predicted or a sufficiently large number to demonstrate the suitability of the method. 'Hold n out' procedures are not appropriate in unsupervised learning and in this case predictions must be based on the 'distances' defined between individual structures or classes of structures. Few quantitative approaches to prediction have been reported in this case.

Although very little literature is available in this area, regression analysis is a more accurate technique than methods of pattern recognition, and it is expected to be the preferred approach in applications where quantitative structure - property correlation or property prediction is the primary objective. However, where suitable property data is not available for regression, classification methods could be useful for giving more approximate estimates of activity.

#### 2.5.4 Computational Considerations

The regression analysis and classification methods developed in the present investigation, like most other applications reported up to the present, have been designed for experimental purposes only and would not necessarily be the most efficient approaches in an operational system. The particular problems arising in obtaining viable methods will depend on the purpose for which the methods are con-

sidered and the required scale of application.

Of the two approaches, regression analysis is computationally more straightforward and a wide range of standard statistical procedures is available in this case for implementing methods. In recent years program packages have been developed which allow the relevance of a large number of explanatory variables to be examined in a variety of different ways. In addition to providing details of the numerical solution to the regression equations the packages usually provide other relevant information about the data at the users request.

Similar computational aids are not widely available in the classification case because of the numerous approaches possible in this case. Some program packages have been developed which allow a wide choice of association measures and clustering methods to be applied,<sup>192-196</sup> although most of the algorithms reported to date have been concerned largely with the implementation of clustering techniques, and displaying clusters geometrically.

The demand for large scale applications is probably higher in the classification case, and devising methods which are suitable for this type of application presents a considerable problem. The processing and storage requirements in large scale operations will vary with the particular techniques involved. Clustering techniques, for example, as seen earlier vary considerably in their demands on computer storage and time. For most methods which require the computation of the full similarity matrix the time is roughly proportional to  $mn^2$  (where  $n$  is the number of objects and  $m$  the number of characters), so that increasing the number of objects has a greater effect than increasing the number of descriptors. If methods could be developed which require only part of the similarity matrix this would



lead to an important saving in time and it would also save on storage, since the space taken up by the similarity matrix usually greatly exceeds the space taken up by the original data matrix, particularly in larger data sets. Much of the information contained in a large similarity matrix is not required by the clustering algorithm, and in recent years several procedures have been proposed for cutting down on the amount of redundant information generated<sup>192,193,197,200</sup> However, using these it is often difficult to decide 'a priori' which similarities are required and which are not, and the methods usually involve making a number of approximations. There has so far been very little practical experience of these methods,<sup>197,198,220-205</sup> and it is expected to be some time before the large scale problems are fully worked out. In time the situation will be helped as faster and larger machines become available.

### CHAPTER 3

A Method for the Automatic Classification of Chemical Structures

### 3.1. Introduction

#### 3.1.1. The Basic Approach

The following section deals with the development of a new approach to the automatic classification of chemical structure data. It is based on an unsupervised, hierarchical clustering technique which has been widely applied in the biological sciences and related disciplines, but infrequently used in chemical applications. The methods developed are the first automatic classification procedures applied to chemical structures to employ structural attributes based solely on the two-dimensional structure diagram, and derived automatically from a machine readable representation of the structure diagram.

As the successful outcome of the classification process depends on making suitable choices at several more or less independent stages, the study has concentrated on examining some of the alternatives possible at these stages, and their effects on the classification result in some small scale applications. For reasons discussed previously, the methods developed concentrate on the problems of structure representation and structure comparison, and a wide range of alternatives is considered at each of these stages.

The choice of suitable structure representation itself involves a number of separate issues, each of which is examined in turn. Thus, the question of suitable structural characteristics is considered by examining a range of atom and bond-centred fragments. How best these should be represented internally is also considered, by comparing a variety of different numerical representations.

Another stage involved is deciding on the relative importance of different structural features, and this question is considered during the comparison of association coefficients, when a number of probabilistic measures are examined.

Structural attributes are recognised and assigned automatically by computer, and thus the algorithm developed could be applied without modification to any group of structures. The structures are clustered by the single-linkage method,<sup>46</sup> which is the simplest of the hierarchical clustering techniques. The various properties of the method and its advantages over other hierarchical techniques have been discussed in Chapter 1. Using this approach, each similarity or dissimilarity coefficient must be examined at least once, which means that for a group of  $N$  structures the method has a time dependence of at least order  $N^2$ . An algorithm<sup>206</sup> has recently been developed which reduces this requirement, and makes it possible to deal with of the order  $10^3$  to  $10^4$  structures by this method.

### 3.1.2. Evaluation Problems

One of the difficulties in investigations of this nature is obtaining suitable data to test the validity of methods. Because there are no standard evaluation procedures available, ideally the structures used should enable comparisons to be made with other methods, and they should also have known physicochemical or biological properties for correlation studies. The classifications could be evaluated on the basis of their retrieval effectiveness, and this approach has been considered in investigations into the use of hierarchical clustering techniques in document-based information systems.<sup>1,2</sup> Several procedures have been proposed for

measuring classification effectiveness in terms of precision and recall data, but to use this method effectively it is necessary to obtain experimental systems, which are suitably representative of an operational environment. Obtaining the appropriate conditions would necessarily involve some user interaction, together with access to working systems, and this type of collaboration is often difficult and time-consuming. A more accessible method of evaluation is to use the relations within the data, in this case structure-property relationships, to estimate the predictive power of methods. The possible chemical interpretation of the classifications obtained in this study also provides a basis for method evaluation on a qualitative level.

### 3.1.3 Feasibility Study

Before suitable data sets satisfying the above requirements were extracted from the literature, a preliminary study was carried out on a number of small random data samples taken from the Chemical Abstracts Service Registry File, and search output from the Sheffield Substructure Search System.<sup>17,207</sup> The investigations used simple binary representation of structures, based on augmented atom descriptions,<sup>208</sup> and structure comparisons were based on the number of fragments common to each structure pair. Details of the fragment definition, and the processes involved in deriving the classifications are given in the following section.

The resulting classifications were displayed as dendrograms and these clearly showed that the method had been successful in clustering together structures of a similar chemical type, where present. For example, Figure 1 gives the classification obtained for one of the CAS data samples, in which the steroid structures

present have been clearly identified.

In view of the very simple conditions used, and the crude measure of association, which only takes into account the information shared by each structure pair, these results were encouraging and well demonstrated the potential of the method.

#### 3.1.4. Data Sets

The investigations following these initial attempts were carried out on a number of data samples recently considered by other investigators for automatic classification or structure-property modelling. The main samples used were a group of 20 naturally occurring amino acids, 39 structurally diverse local anaesthetics, and a group of 79 synthetic penicillin structures. Details of these and the available property data are given in Appendix 1.

The amino acids were taken from an investigation by Sneath<sup>10</sup> and were useful because they allowed comparisons to be made with an alternative classification approach in which structures are represented by a combination of manually derived structural attributes, and physical, chemical and biological properties.

The local anaesthetics, taken from the work of Agin et al,<sup>209</sup> were another useful group, as they tested the ability of the methods developed here to handle diverse structural types. They also allowed comparisons to be made with Agin's semiempirical, quantum-chemical approach to the structure-property problem.

Finally, the penicillin structures were taken from a study by Bird and Marshall,<sup>210</sup> who looked at the relationship between the serum binding properties of penicillins and the hydrophobic character of

their side-chains, expressed in terms of Hansch substituent values. These therefore enabled useful comparisons to be made with this widely used semiempirical approach to structure-property correlation.

### 3.2 A Comparison of Some Alternative Numerical Representations of Substructures

Detailed investigations of association measures and substructural fragments using the above data samples were preceded by a comparison of a number of different numerical representations, in which different amounts of detail are recorded about each substructure. A simple binary representation, recording the presence or absence of each fragment type is compared with a more detailed representation indicating the different occurrences of each fragment type, and this, in turn, is compared with a representation based on multiple fragment occurrences, which does not distinguish between the equivalent structural features arising in chains and non-aromatic ring systems.

Investigations were carried out on the amino acid and anaesthetic structures, and three simple coefficients of association were used for structure comparison. These coefficients are preferable to the very simple measure considered previously, because they take into account the number of unshared fragments in each structure pair, and they are also normalised. Negative score agreements are also taken into account in some cases. The numerous coefficients of association proposed for binary strings have been discussed in Chapter 1, and the particular coefficients considered here are described below. A more detailed discussion of the coefficients used is also given in Section 3.3.

### 3.2.1 Method

#### 3.2.1(a) Substructures

Each structure used in the analysis is coded as a redundant connection table<sup>15</sup> for input to the computer. Bonds were divided into five types for coding, namely, single chain, single ring, double chain, double ring, and aromatic ring bonds. Tautomeric bonds were not represented as such, but were reduced to one of the five possibilities listed above.

Similarity coefficients (SCs) and dissimilarity coefficients (DCs) were based on the presence or absence of augmented atoms<sup>208</sup> in the structure. These are centred on each atom of each structure, and consist of the central atom, the bonds it forms and the atoms to which it is bonded, excluding hydrogen atoms and bonds to hydrogen. Figure 2 shows the augmented atom fragments occurring in the amino acids, asparagine and glutamine.

#### 3.2.1(b) Numerical Representations

The following three descriptions in terms of augmented atoms were used,

- (i) The presence or absence of an augmented atom type in a structure was noted. The second and subsequent occurrences of the same augmented atom type in the same structure were ignored.
- (ii) Each occurrence of an augmented atom type in a structure was noted. A suitable set of attributes was selected to cover the different occurrences in each structure and multiple occurrences of the same augmented atom type were allowed for in the calculation of SCs and DCs by an additive coding method.<sup>10</sup>



(iii) Multiple occurrences of augmented atom types were treated as in (ii) but only three bond types namely alternating ring bonds, single bonds and double bonds were discriminated. Thus, in contrast to (i) and (ii) ring and chain bonds were not differentiated in the case of double and single bonds.

All three representations were considered in the set of amino acids, and structure representations (i) and (ii) in the case of the anaesthetics.

The first stage in the calculation was to analyse each set of structures and note all of the augmented atom types which occurred. This gave a list of all the attributes upon which the calculation of SCs and DCs were based. Next, a description of each structure in terms of the set of attributes was formed and stored in a bit vector. Each pair of bit vectors was then compared to calculate the SC or DC between the corresponding pair of structures. The particular additive coding method considered involves some logical redundancy, as the attributes selected for each augmented atom type are not mutually exclusive. This approach has been discussed in detail in Chapter 1.

### 3.2.1(c) Structure Comparison

For each pair of structures the attributes were divided into four groups containing a, b, c and d attributes, where 'a', 'b', 'c' and 'd' are the entries in a 2 x 2 contingency table, i.e. '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

to which this SC or DC refers.

The three coefficients<sup>10, 46</sup> used were: -

1	Dice's SC	$\frac{2a}{2a+b+c}$
2	$\emptyset$	$\frac{(ad-bc)}{[(a+b)(a+c)(d+b)(d+c)]^{\frac{1}{2}}}$
3	Sneath's DC	$\frac{b+c}{a+b+c+d}$

The matrix of SCs obtained for the amino acids using  $\emptyset$  and structural representation (ii) is shown in Table 1.

### 3.2.1(d) Clustering

The structures were finally classified by the single-linkage method using an agglomerative algorithm originated by van Rijsbergen.<sup>74</sup> Dendrogram representations were derived manually from the cluster listings produced by the clustering algorithm. The particular clustering procedure implemented by the algorithm is a modification of the single-linkage clustering method described by Sneath.<sup>34</sup> Each level of association arising in the matrix is examined in turn and initially structures related with the highest possible SC or lowest DC values are clustered. Successively lower levels of association are then examined. The criterion of admission of a new structure into a cluster is that the new member should be associated with at least one of the members of the existing cluster at the given SC or DC value. Similarly, for the union of two clusters at least one member from each cluster should be associated at the given level of association. The basic difference between this approach and Sneath's method is that the levels at which clusters are formed are based on the SC and DC values arising within the group and are not arbitrarily set at equally spaced intervals. A simple example is given in Figure 3, which

shows the initial clusters formed for amino acids asparagine, glutamine, arginine and lysine using the SCs derived from  $\emptyset$  and structure representation (ii). The SC levels at which clusters are formed are underlined. Moving down the hierarchy asparagine and glutamine are the first structures to form a cluster at level 0.94. At the next highest association level of 0.77 arginine and lysine also form a separate cluster, and the initial cluster remains unchanged. Finally the two clusters join at level 0.74 due to an association between lysine and glutamine at this level. From a computational point of view one of the advantages of the single linkage method over other hierarchical clustering methods is that it is not necessary to construct a new association matrix at each new level examined, as clusters are always formed on the basis of associations between individual members.

### 3.2.2 Results and Discussion

#### 3.2.2(a) Predictive Performance

The relative usefulness of the structural representations were assessed by simulating the 'predictive' use of the SCs and DC, and the classifications derived from them. Unfortunately the gross approximations of the method and the dependency of predicted property values on observed property values prevents the application of significance tests, such as chi-square, to measure the significance of the differences between the predictions obtained by alternative methods. Apart from showing possible predictive value, the predictions are therefore only intended to provide rough guidelines to the usefulness of methods, and to illustrate possible trends in cases where different techniques are compared.

The dendrogram representations of the classifications serve a similar purpose.

A different property was considered for each of the data sets. The properties available in the case of the anaesthetics were Log (MBC) i.e. minimum blocking concentration values used by Agin et al.<sup>209</sup> In the amino acids pI values<sup>211</sup> were used.

(i) The Similarity and Dissimilarity Coefficients

It was assumed that the property of each structure in turn was not known and its property was set equal to that of the structure with which it was most similar according to the values of the SC or DC in question. The average value of the difference between observed and predicted property values was then calculated. Where a structure's highest similarity was with two or more others the average of the property values was taken.

The results obtained for the amino acids are given in the lower entries in the cells in Table 2. The best result was obtained using Dice's SC and  $\emptyset$ , and structural representation (ii). These both gave average deviations of 0.43 pI units between observed and predicted values. It is instructive to compare the average deviation obtained with two other values which could be obtained under other circumstances. A high value would arise if it were not possible to form classes of the amino acids from the 20 studied. In this case the predicted value for any acid would be the average pI value taken over the other 19 structures. The average deviation in this case is 1.08 pI units. The smallest possible value which could be expected would occur if each acid had its highest SC or lowest DC with the acid which also had the nearest pI value. In this case the average deviation between observed and

predicted values would be 0.26 pI units. Dice's coefficient is thus close to the smallest possible value of 0.26 for this set of structures and is very much less than the values of 1.08 which would have resulted if no resolution of structures had been obtained.

An examination of Table 2 shows the level of prediction to improve as the structural representation becomes more detailed, and the result obtained using structure representation (ii) is very much better than it is with representation (i). Using these two representations a similar improvement is observed in the local anaesthetics, as shown in Table 3. In this case structure representation (ii) and Sneath's DC give the lowest average deviation of 0.84 log (MBC) units. Compared with the amino acids this value is not quite as good when viewed against the smallest deviation possible in this case of 0.09. However, it is still a considerable improvement over the value of 1.69 which would be obtained if no resolution of structures had occurred. The mean deviations obtained using structure representation (i) are much closer to this value.

(ii) The Classifications

The classifications were tested in a similar manner to the association coefficients but the 'predicted' value of a structure in this case was taken to be the average property value of the cluster which it joined. The best predictions were again obtained using structure representation (ii). The remaining entries in Table 2 give the results obtained for the amino acids. As before,  $\emptyset$  and Dice's SC performed best, giving an average deviation of

0.39 pI units. Dice's SC also gives the lowest deviation in the anaesthetics (mean deviation 1.16 log (MBC) units), and the results in this case are shown in the upper entries of Table 3.

### 3.2.2(b) Structural Arrangements

In all cases the clusters obtained are sensible from a general, qualitative chemical point of view, and the examples given in Figures 4 to 8 show that there is a gradual improvement in the resolution as the level of detail included in the structure representation is increased. These results are in agreement with the different levels of prediction obtained.

The dendrogram representations show that structure representation (i) is not sufficiently powerful to distinguish between some of the more closely related amino acid structures, such as the two acidic amino acids, aspartic and glutamic acids, and the amides asparagine and glutamine. These pairs are differentiated at the two higher levels of description. Similarly in the anaesthetics, structure representation (i) has successfully identified the important structural types, but it is unable to distinguish between some of the more closely related structures present e.g. the group of normal alcohols, and other structures involving the same substructural components, e.g. phenol and hydroquinone, and quinoline and phenanthroline. These different groups are again resolved at the most detailed level of description.

The ability of the different structural representations to separate cyclic and acyclic derivatives varies in each sample. In the closely related group there is a gradual improvement in the separation as the level of description becomes more detailed.

In the structurally diverse group, because of other, wider differences arising structure representation (i) gives a more satisfactory separation here, and the two representations considered (i.e. (i) and (ii)) give roughly equal performances.

The classifications for the amino acids using structure representation (ii) correspond closely to those described by Sneath<sup>10</sup> and Meister.<sup>211</sup> Thus, the structures show a broad breakdown into cyclic and acyclic classes and two clearly defined clusters are formed between the two carboxyl acids and the two amides. Lysine and arginine, which contain two  $-NH_2$  groups also form a separate cluster. The acyclic hydroxy amino acids, serine and threonine do not cluster initially, but these join the same cluster at different levels. Sneath's DC values, reclassified by the single-linkage method also give a very similar result (Figure 9). This close agreement with Sneath's results is encouraging, and it illustrates the usefulness of systematically derived structural descriptors, compared with physicochemical parameters and structural attributes preselected on a chemical basis.

Graphs of observed against predicted property values were plotted to illustrate the strength of the relationship between the structural classifications and the property in question, and examples of the type of agreement obtained in each sample are given in Figures 9A and 9B. In the amino acids sample the scatter of points clearly shows the ability of the method to discriminate between the three main groups present i.e. the two strongly acidic structures, the two strongly basic structures, and the remaining amino acids which have almost neutral pI

values. The results in the anaesthetics are not so easily interpreted, as there is not such an obvious relationship in this sample between structure and activity. The predictions obtained in this case are discussed more fully in the following section.

In the comparison of approaches the study showed that the mean deviations between observed and 'predicted' property values are a useful estimate of the predictivity of methods, when they are compared with the best possible result which could be obtained under the given circumstances. However this measure does not take into account the numerical range of property values covered by the data set, and it is not the most suitable quantity to consider when making comparisons across different data samples. Ideally, property deviations should be looked at in relation to the observed property range, and a more reliable estimate of the agreement between observed and 'predicted' properties can be obtained by taking the sum of squares of the differences between observed and 'predicted' values as a ratio of the sample variance. This more frequently applied statistic was therefore used in subsequent investigations of association measures and substructural fragments. The most satisfactory numerical representation here, noting fragment occurrences, was also used in these studies and multiple occurrences were recorded either in a series of two-state attributes, as above, or in a single, multi-state or quantitative attribute, depending on the association measure used.

### 3.3 An Evaluation of Some Different Measures of Resemblance

The large numbers of similarity and dissimilarity measures available for use in automatic classification applications were



discussed in detail in Chapter 1. As very few comparative studies have yet been carried out in this area, the essential differences between coefficients are not fully understood, and in most studies this presents the user with a serious problem of choice.

As seen in Chapter 1 many of the available coefficients require a qualitative, binary representation of the data, others need a quantitative description, whilst others can be applied equally well in either case. A variety of these coefficients have been examined, with the exception of correlation coefficients which are now thought to be of questionable value in automatic classification applications of this nature.

The investigations look at a number of simple matching coefficients and compare these with a distance measure and a number of probabilistic measures. The simple matching coefficients are an important group to consider because of their widespread application in a wide range of disciplines, and their simplicity both from a conceptual and computational point of view, in comparison with more sophisticated quantitative measures. Their main limitation is that they require a binary representation of the data. As discussed in Chapter 1, the essential differences between matching coefficients lies in their treatment of matched and unmatched pairs of values in two individuals, and in their treatment of negative matches (see also section 3.2). These differences are examined in a number of different formulations to see whether or not they have a significant effect on the classification result, and in particular whether the inclusion of negative matches is important.

Distance measures are another group of coefficients which

have been widely used in automatic classification applications, and they are a particularly interesting group in the present investigation because of their recent widespread consideration in applications of pattern recognition techniques to chemical structure data.<sup>20,54,55,153</sup> The value of these coefficients is that they can handle quantitative descriptions of the data, and although other coefficients have been proposed which can do this, distance measures are more appealing from a conceptual point of view and are usually more straightforward computationally than other forms of quantitative measure.

Probabilistic coefficients are another useful group to consider, as these have so far been very infrequently applied in numerical classification, due to the large amounts of computation usually involved. To date no applications of this type of coefficient have been reported in the chemical literature. As discussed in Chapter 1, one of the more usual arguments put forward in favour of character weighting is that infrequently occurring characteristics are more discriminating and should be more heavily weighted than frequently occurring states. A number of probabilistic coefficients have been therefore examined based on this premise.

For a meaningful comparative study of the above coefficients and of the classifications produced from them, the same structural representation was used throughout, i.e., the same substructure was considered in each case, and one of two equivalent numerical representations of this was employed depending on the type of resemblance measure in question.

Similar coefficient performances were given by the amino acids, local anaesthetics and penicillins, and only the results

obtained for the local anaesthetic structures have been reported below.

### 3.3.1 Method

#### 3.3.1(a) The Structure Representations

As in the investigation of structural representations, the structure of each anaesthetic was described as a redundant connection table, and this was used to obtain a set of augmented atoms upon which measures of similarity and dissimilarity were based. The anaesthetics were first analysed to identify the different augmented atoms arising, 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 of structures 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.<sup>10</sup> This corresponds to structure representation (ii) considered in section 3.2.
- (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.

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

### 3.3.1(b) The Similarity and Dissimilarity Coefficients

#### (i) Association/Matching Coefficients

The three simple coefficients of association examined were used in the previous investigation of numerical representations, and details of their formulation are given in Section 3.2. For completeness these are listed again below, and the numbering of expressions corresponds to the numbers used in Table 4.

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

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

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

These three coefficients were applied to structure representation (i)

Both Dice's SC and Sneath's DC are normalised within the range 0 and 1, whereas  $\emptyset$  values lie within the range -1 to +1. The important difference between Dice's SC and the remaining two is that it does not take into account negative score agreements between pairs of structures. In this SC, matching pairs of binary values carry twice the weight of disagreeing pairs, which means its magnitude is greater than other similar coefficients based on positive matches, in which matched and unmatched pairs of values are weighted equally, e.g. Jacchard's coefficient. The dissimilarity coefficient defined by Sneath is the complement of Sokal and Michener's simple matching coefficient<sup>36</sup> expressed as a percentage, and using this, matched and unmatched pairs of binary values are weighted equally. Matched and unmatched pairs of values are also

equally weighted in the  $\emptyset$  coefficient, and in this case agreeing pairs of values, including negative agreements, are balanced against disagreeing pairs of values in the numerator.  $\emptyset$ , which is also frequently referred to as the four-point correlation coefficient is a measure often considered in statistics because of its relation to  $\chi^2$ . In this particular application however the arrangement of the data upon which the measure is based cannot be compared with the conventional 2x2 contingency tables used for tests of independence in statistics, and it is doubtful whether any meaning can be attached to such a test in this case.

(ii) Distance Coefficients

The distance measures considered are based on the simple Euclidean distance measure proposed by Sokal.<sup>90</sup> Thus, the number of occurrences of each augmented atom type in a structure is regarded as a metric quantity, and the similarity between pairs of structures is expressed in terms of their distance apart in an n-dimensional space, where the coordinates represent the n different augmented atom types arising in the total set of structures. Distances are first computed between individual augmented atoms and these are then summed over all fragment types to gain an overall measure of dissimilarity between structure pairs, as follows,

$$\delta_{jk}^2 = \sum_{i=1}^n (X_{ij} - X_{ik})^2$$

where  $X_{ij}$  is the number of occurrences in the jth structure of the augmented atom fragment defined by attribute i. As in Sokal's basic formulation the squared distance is used as the measure of dissimilarity in order to avoid square root terms on the right hand side of the expression.

Problems due to differences in scale between the attributes involved in the distance expression do not arise in this case, but it is possible that differences in the range of frequency values arising for each fragment type could lead to an unsatisfactory result. To compensate for possible distortions, therefore a second distance measure is computed after first standardising frequency values so that each attribute possesses a zero mean and unit variance (coefficient 4a in Table 4).

The above distance measures were applied to structure representation (ii)'.

(iii) Probabilistic Coefficients

The three coefficients considered in this category are based on the probabilistic similarity index proposed by Goodall.<sup>77</sup> During the comparison of pairs of structures, each attribute is considered in turn and the weight attached to the particular pair of values arising for that attribute is calculated from the likelihood of that pair of values, or a more 'similar' pair of values arising according to Goodall's definition of similarity for individual attributes. The definition of similarity depends on the type of attribute in question. Thus, in the case of the qualitative two-state binary attributes in structure representation (i)' pairs of differing binary states are regarded as being equally dissimilar, and the similarity between agreeing positive and negative binary scores is based on the probability of the particular pair of values arising. The less probable the match in question the more similar the structure pairs are said to be with respect to this attribute. The binary attributes considered here are a special case of the qualitative attributes discussed by

Goodall, and have been treated in a similar manner. The similarity (S) between attribute values in this case may be summarised as follows: \*

For binary states i and j

$$i \neq j \quad \supset \quad S_{ij} = 0$$

$$i = j \quad \supset \quad S_{ij} > 0$$

and the probability term associated with values i and j is

$$P_{ij} = \sum_{k \in Q} P_k^2, \quad i = j$$

where  $Q = \{k: (p_k \leq p_i)\}$

ie. the set of all k ( $1 \leq k \leq n$  being understood) such that

$$p_k \leq p_i ;$$

$$\text{and} \quad P_{ij} = 1, \quad i \neq j$$

The similarity between attribute values i and j is then defined as follows

$$S_{ij} = 1 - P_{ij}$$

The definition of similarity is more complicated in the case of the quantitative augmented atom descriptions given in structure representation (ii)', since the magnitude of attribute values must also be taken into account here. The different values which an attribute may take, representing the frequency of occurrence of each augmented atom in a structure, are now treated as metric quantities, and Goodall's definition of similarity for ordered, metrical attributes is used. Thus, in two structures, agreeing attribute values are considered more similar than values which differ, and those with a small difference are considered more similar than those with a larger

\* the following terms from symbolic logic are used to simplify the presentation of expressions,  
 $\supset$ (implies),  $\varepsilon$ (is included in),  $\wedge$ (and),  $\vee$ (inclusive or),  $:$ (such that)

difference. If pairs of values should differ by the same amount these are resolved in the same way as qualitative attributes, described above, i.e. by taking into account the likelihood of each pair of values arising. Thus, for a given attribute which takes on frequency values  $V_i$  and  $V_j$  in one pair of structures and values  $V_k$  and  $V_l$  in another,

$$|v_i - v_j| < |v_k - v_l| \quad \supset \quad s_{ij} > s_{kl}$$

$$(|v_i - v_j| = |v_k - v_l|) \wedge \left( \sum_{t=1}^j p_i < \sum_{t=k}^l p_i \right) \quad \supset \quad s_{ij} > s_{kl}$$

$$(|v_i - v_j| = |v_k - v_l|) \wedge \left( \sum_{t=1}^j p_i = \sum_{t=k}^l p_i \right) \quad \supset \quad s_{ij} = s_{kl}$$

and the probability term associated with pairs  $ij$  is

$$P_{ij} = \sum_{k \in Q} \{ p_k^2 + 2 \sum_{t \in T_k} p_k p_t \}$$

where  $Q = \{k: [(i \neq j) \vee (p_k \leq p_i)] \wedge [1 \leq k \leq n]\}$

and

$$T_k = \{t: [(|v_t - v_k| < |v_i - v_j|) \vee [(|v_t - v_k| = |v_i - v_j|) \wedge$$

$$\left( \sum_{u=1}^j p_u \geq \sum_{u=k}^t p_u \right)]\} \wedge |k < t \leq n| \}$$

The third probabilistic measure considered was also based on structure representation (ii)', except in this case a modification of Goodall's definition of similarity for quantitative attributes was used, where the identification of pairs of frequency values which are more similar than the particular pair of values in question is based on frequency values alone. Thus, in the above expression, the attribute pairs  $k, t$  which qualify for inclusion in the probability expression derived for values  $i, j$  are those which satisfy the following condition,



$$T_k = \{t: |v_t - v_k| \leq |v_i - v_j|\}$$

In practice, the probabilities of different attribute values are not exactly known and must be estimated from the sample of individuals in question on the basis of observed frequency values. Also, to reduce the amount of computation involved, each attribute in a structure is assumed to be independent of the others.

Having obtained probability terms for individual attributes in the above manner, an overall measure of similarity between pairs of structures is then obtained by summing the appropriate probability terms over attributes in the following way,

$$\sum_{x=1}^n -\log P_{x,1,2}$$

where  $P_{x,1,2}$  is the similarity term derived for attribute  $x$  in structure pair 1 and 2, and  $n$  is the total number of attributes. Negative logarithmic terms are taken as a measure of similarity in preference to complement values. This is a very much simplified version of the similarity expressions derived by Goodall between individuals, where, following the ordering relations derived for individual attributes, he computes the cumulative probabilities over total sets of attributes to determine the likelihood of the particular pair of attribute sets for the individuals in question or any more similar pair of sets arising.

The above three probability measures appear in Table 4 as coefficients 5, 6 and 7 respectively, and coefficient 7, based on the modified version of Goodall's definition of similarity for ordered, metrical attributes, is referred to in the text, as the 'minimum distance' coefficient, so as not to be confused with coefficient 6.

3.3.2 Coefficient Performance

3.3.2(a) Nearest Neighbours

Following the evaluation procedures used in the investigation of numerical representations of structural attributes, the similarity and dissimilarity coefficients described above were compared by simulating their predictive use. Thus, to obtain a 'predicted' value for the property of each anaesthetic, the structure with which it is most closely associated was used. The observed log (MBC) values given by Agin et al,<sup>209</sup> were considered as before, and where more than one nearest neighbour arose the average log (MBC) value over the set of nearest neighbours was calculated.

For each resemblance coefficient, the sum of the squares of the differences between observed and predicted log (MBC) values, taken as a ratio of the sum of the squares of 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 and both these measures were taken as an indication of the effectiveness of the different coefficients under examination. The property deviations are shown in Table 4. The lowest sum of squares ratio and mean deviation was obtained using the squared distance coefficient, which gave a sum of squares ratio of 0.34 and a mean deviation between observed and estimated property values of 0.79 log (MBC) units. As in Section 3.2 the mean deviations were put into perspective by comparing them with the best possible result which could be obtained for the given set of values, the mean deviation which would have resulted if there had been no resolution of the anaesthetics into classes, and, finally, the mean deviation of observed property values from

their mean value. These quantities are 0.09, 1.69 and 1.65 log (MBC) units respectively. The deviation of 0.79 log (MBC) units produced by the squared distance coefficient is therefore a good improvement on the mean deviation for the total set and the mean deviation assuming a homogeneous group. The worst result was obtained using probabilistic coefficient 6, based on the quantitative frequency descriptions given in structure representation (ii)', where the mean deviation between observed and estimated property values of 1.891 log (MBC) units exceeded both the above values. The sum of squares ratio also exceeded unity in this case. The probabilistic measure based on structure representation (i)' also gave a poor result.

With the exception of the above two probabilistic coefficients, the reduction in the variance as measured by the sum of squares ratio is reasonably good. However, as previously mentioned, it is not possible to evaluate these results from a rigorous statistical point of view, but in view of the method of prediction and the sample size in question it is unlikely that the very close values obtained by the different coefficients are significantly different, with the possible exception of the results given by the squared distance coefficient compared with those obtained using the probabilistic measures 5 and 6.

The extent of the relationship between the property in question and the structural differences as measured by the distance coefficient is shown in Figure 10, which gives a plot of observed log (MBC) values against the predictions simulated on the basis of this DC.

Using each coefficient a number of the anaesthetics are very

well predicted, for example the structural isomers phenyltoloxamine and diphenhydramine. Other structures which have been well predicted in each case are the group of normal alcohols, ranging from n-propanol to n-octanol. Methanol is not included in this series due to the absence in its structure of a methylene group, making its association with ethanol very much weaker than the latter's association with propanol. Using the two distance coefficients and Sneath's DC, each alcohol in the group, except for the terminal members, are equally highly associated with the two alcohols adjacent to it in the series, whereas the similarity coefficients give closest associations with the next highest alcohol present. These different relationships are brought out in the dendrogram representations illustrated in Figures 8, and 11 - 17, and are discussed in the following section, describing the classifications obtained. The first of the above cases is an interesting one as the predicted property value for each alcohol is the value which would be obtained by linear interpolation from the two nearest neighbours in the homologous series. This in fact gives a better prediction than the second case, the reason being that the alcohols are fairly widely distributed through the particular structure set from an activity point of view. Thus, predictions based on the mean of two adjacent values are closer to the observed value of the alcohol in question than are either of the two adjacent values taken independently.

Apart from the above well defined structural types some variations are observed, depending on the resemblance coefficient applied. Quinolone and 8-hydroxyquinolone, for example which are similar both in structure and activity are well

predicted in the case of the simple association coefficients and distance measures, but are poorly predicted by the probabilistic coefficients, with the exception of coefficient 6 which gives a good prediction for quinoline. The reason for this is that the probabilistic factors coming into play with these measures do not necessarily relate the most similar structural types. In the present sample the most similar structural types do not always lie closest together on the scale of activity, for example, the alcohols mentioned above are fairly widely distributed through the group, and therefore this particular property of the probabilistic measures is not necessarily an undesirable one. However, in the present example it has an adverse effect on the result, giving quinoline and 8-hydroxyquinoline closer associations with structures such as 0-phenanthroline and quinine, which are much further apart on the activity scale in question. Using the 'minimum distance' probability coefficient, 8-hydroxyquinoline is most closely associated with phenol, which again results in a poor prediction.

Some of the anaesthetics have been poorly predicted in every case because they do not belong to a distinct chemical group and form no other strong associations, e.g. eserine, dibucaine and quinine. Chloroform which is another structure belonging to this category is an unusual case as it contains a unique set of augmented atom fragments, and it is instructive to see how the different coefficients handle this particular situation. Thus, the simple matching coefficients  $\emptyset$  and Dice's SC, which depend on the number of structural features in common to each structure pair, give negative and zero levels of association respectively between this structure and the remaining structures present, as

these have no features in common. The dissimilarity and distance coefficients on the other hand base similarity on the 'distances' between structures on the smallest number of unshared features. In these cases therefore, chloroform associates with some of the smaller structures in the sample because these structure pairs are the least dissimilar with respect to the particular DC in question. Thus, using Sneath's DC and the squared distance coefficient, chloroform forms a definite association with methanol, despite the fact that these structures have no augmented atoms in common. A similar situation arises with the probabilistic coefficients, as again in these cases similarity is not merely based on a straightforward matching of shared attributes. Using all three coefficients in this category, chloroform associates most closely with the saturated bridged ring system antipyrine, which although close in activity, is structurally very dissimilar. These examples show how coefficients which do not rely on a straight matching of common structural features may give rise to groupings which are not the most similar chemically, and this could be an important consideration in applications concerned primarily with structure organisation for retrieval.

### 3.3.2(b) Classifications

As before, classifications were assessed on the basis of their predictive value, and on the chemical significance of clusters.

#### (i) Simulated Predictions

In this case the 'predicted' property value for each anaesthetic is taken to be the average log (MBC) value for the cluster which it joins. The predictions are given in Table 4 and

they show how the performances in this case follow closely the results obtained using nearest neighbours. Thus, the probabilistic measures again perform poorly, particularly coefficients 5 and 6, and the mean deviation of 1.973 log (MBC) units using coefficient 6 again exceeds the sample mean deviation and the deviation resulting if no resolution of structures into classes had taken place. As before the sum of squares ratio also exceeded unity in this case. Dice's SC gave the best sum of squares ratio, and the mean deviation between observed and predicted property values was again lowest when using the squared distance coefficient. With the exception of probability coefficient 6, predictions obtained from nearest neighbours are slightly better than those given by the classifications, which is not an unreasonable result, in view of the information loss accompanying the transformation from an association matrix to a dendrogram and the diverse structural types present. Again, the different levels of prediction lie reasonably close together, with the exception of the very good prediction obtained using the squared distance coefficient and the poor predictions obtained using probability coefficients 5 and 6. However, as discussed earlier the method of prediction used and the dependency of estimated property values on observed values prevents the application of statistical tests to compare the different levels of prediction, and it is therefore not possible to say with certainty whether any of the values are significantly different from each other. On the basis of the present evidence it would therefore be unwise to draw any firm conclusions on the performances of the different coefficients, and the only possible way of obtaining meaningful comparisons with this type of investigation would

be to attempt to show up possible trends on an empirical basis by applying the coefficients to a much wider range of structures.

(ii) Structural Arrangements

From the previous discussion it is clear that the levels of prediction obtained by the different resemblance measures are of limited value in comparative studies. However, comparing the overall structural arrangements obtained by each measure it is interesting to note that the coefficients giving the best predictions also give the sharpest resolution of structures from a chemical point of view. Thus, the simple association coefficients, the squared distance measure and the 'minimum distance' probability coefficient, which give reasonably good levels of prediction, also show some well defined chemical groups, and a very clear breakdown into cyclic and acyclic classes at an early stage in the classification i.e. at a low level of association. Using the squared distance coefficient, for example, there is a very definite split between cyclic and acyclic structural types, with the exception of antipyrine which associates with the acyclic group, and of eserine, diphenhydramine, phenyltoloxamine, caramiphen and quinine, which have cyclic and acyclic components of comparable size. This second group forms no strong associations with the remainder of the set of structures and except for the two structural isomers, phenyltoloxamine and diphenhydramine, which form a separate cluster, these join classes by chaining at a much lower level of similarity. With a few minor variations a similar pattern is followed by the other coefficients listed above. Probability coefficients 5 and 6 on the other hand, which give worse predictions, give a poorer resolution of structures in comparison and do



not show a clear division into predominantly cyclic and acyclic classes. Similarly, the classification using the standardised distance coefficient is less satisfactory from a qualitative chemical viewpoint, and this also gives a poor prediction.

Some of the coefficients have given rise to very similar arrangements, although none of the measures considered are jointly monotonic. For example,  $\emptyset$  and Dice's SC give almost identical classifications, and Sneath's DC and the squared distance coefficient also give close results. The standardisation of frequency values in the case of the distance measure has had an adverse effect on the result, and the classification obtained in this case bears little resemblance to the arrangement given by the non-standardised measure. The classification using the 'minimum distance' probability coefficient also bears little resemblance to arrangements given by the other two probability measures. In this case a much clearer resolution of structures is obtained and the arrangements correspond more closely to those given by Dice's SC and  $\emptyset$ .

In addition to the above very close agreements, all the coefficients showing a broad breakdown into cyclic and acyclic classes give reasonably similar arrangements overall. In each case the smaller group of acyclic structures as mentioned earlier reveals a well defined cluster of normal alcohols and the slightly different arrangements produced by similarity and dissimilarity coefficients, due to the different relationships these different measures give between adjacent members of the group, are clearly indicated. Using Sneath's DC and the squared distance coefficient methanol and isopropanol join the alcohol cluster

and these fuse at a slightly lower level of association, due in both cases to an association with ethanol. Using  $\emptyset$  and Dice's SC isopropanol also joins the alcohol cluster but in these cases methanol is completely dissociated from the group. This is because it has no augmented atom fragments in common with the rest of the group, which is an important requirement with this type of similarity measure, as discussed earlier.

The larger group of cyclic structures first shows a broad breakdown according to the size of the ring system present, with structures incorporating larger acyclic components separating from structures 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. Thus, the latter group reveals a well-defined cluster of simple benzene derivatives, consisting of toluene, phenol, benzyl alcohol and hydroquinone. A similar arrangement is also given by the standardised distance measure, except that in this case phenol and hydroquinone initially form a separate cluster and join toluene and benzyl alcohol after these two have joined the alcohol cluster. In contrast, none of the probabilistic measures, including the 'minimum distance' probability coefficient show a definite relationship between these different benzene derivatives, and these are dispersed through the group.

The nitrogen containing heterocyclic ring derivatives present, involving small acyclic components also form some well defined clusters. Some of the associations formed in this case and the differences arising with each similarity and dissimilarity measure have been discussed in the previous section. Using Sneath's DC,

the squared distance coefficients,  $\emptyset$  and Dice's DC, quinoline and 8-hydroxyquinoline form a separate cluster at a high level. 0-phenanthroline is also associated with this cluster when  $\emptyset$  and Dice's SC are applied, and pyridine in the case of the standardised distance coefficient. However, the remaining nitrogen heterocycles arising are not associated with these structures. In the case of benzimidazole this is because the five-membered heterocyclic ring present is classed as a saturated ring and is coded with localised ring bonds, giving it a much lower association with the other nitrogen heterocyclic derivatives present, which are classed as unsaturated systems. In all other cases, it is because of the more powerful influence of larger acyclic components. A similar scattering of N-heterocyclic structures occurs using the probabilistic measures, although the above arguments do not necessarily apply in this case. As seen earlier, these measures have the ability to bring together quite dissimilar structural types, and in this case, not even quinolone and 8-hydroxyquinoline form a separate cluster initially. Probability coefficient 5 initially forms separate clusters between quinolone and 0-phenanthroline and between 8-hydroxyquinoline and quinine, and these pairs eventually join to form a single cluster at a lower association level. Using probability coefficient 6, quinoline does in fact associate with 8-hydroxyquinoline first, but the latter structure again forms a stronger association with quinine, resulting in a separate cluster between these two structures which is eventually joined by quinoline. The 'minimum distance' probability coefficient gives rise to a separate cluster between quinoline and 0-phenanthroline, and in this case 8-hydroxyquinoline is more closely related to the phenolic derivatives present, initially forming a cluster with 2-naphthol and phenol.

Within the cyclic group containing larger acyclic components the associations vary from coefficient to coefficient, but in general the main clusters are formed between the dialkylamino ethyl ester derivatives of benzoic acid and the dialkylamino derivatives of acetanilide. The squared distance coefficient for example gives a well defined cluster between procaine, tetracaine and xylocaine. Dibucaine and caramiphen, which also have similar acyclic substituents are not included in this group due to the different types of ring systems present. However, using the two similarity measures,  $\emptyset$  and Dice's SC, a cluster is formed between procaine, tetracaine and caramiphen despite the fact that caramiphen contains an additional 5-membered saturated ring system. In this case xylocaine chains at a lower association level. A closer examination of these structures in Appendix 1 shows that the side chain in procaine is more closely related to the side chain appearing in caramiphen than that arising in xylocaine. As a result, the number of shared features between procaine and caramiphen is greater than the number shared between procaine and xylocaine, and therefore when the above two similarity measures are applied a closer association is formed between the first pair. Using the squared distance measure, however, the reverse situation arises, as the presence of the 5-membered ring now increases the number of unshared features present compared with those arising between procaine and xylocaine. Another pair of structures which have side chains similar to those arising in this group are phenyltoloxamine and diphenhydramine. These two structural isomers form a clearly defined cluster in every case but they are dissociated from the rest of the structures in this group because of the diphenylmethane ring configuration.

Although such clearly defined chemical groups do not arise with the probabilistic measures, particularly when coefficients 5 and 6 are applied, some of the more closely related structural types are still clustered satisfactorily in these cases. For example, the above named structural isomers still form a separate cluster in each case, and the normal alcohols are also clearly identified.

### 3.4 Choosing Suitable Substructures

#### Introduction

Obtaining meaningful descriptions of the original data, which convey as much of the original information as possible and are at the same time suitable for representation in numerical form, involves, as seen in Chapter 1, a number of separate and very important issues, and this is probably the most critical stage of the classification process. The particular problems arising in chemical applications have been outlined in Chapter 2. Using representations based on the structure diagram one of the main difficulties, discussed in this earlier chapter, is choosing fragments of a suitable size which will bring out all the features of possible chemical relevance. Thus, larger fragments provide more detail on ring systems and ring substitution patterns whereas smaller fragments have the advantage of being able to identify important functional groups, often masked by larger substructures. Another difficulty arising, particularly in the case of sub-

structures derived automatically from a connection table representation of the structure diagram, is the degree of overlap between descriptors. This becomes more serious as the fragment size increases, and with larger fragments the variety of substructures also increases. One possible solution to the masking problem with large substructures would be to include a number of the smaller definitions with these. However this would lead to further increases in the amount of redundant information held.

In view of the above considerations a number of different fragment definitions have been examined, and some of these have been combined in two additional representations, to give some indication of the effects of this type of multilevel description on the classification result.

Investigations so far have shown that it is difficult to draw very definite conclusions on the suitability of different techniques, because of the difficulties of measuring statistical differences between method performances, and comparing them on a quantitative basis. The measures considered are useful for illustrating trends in the data and if similar trends are observed in different samples then these are more likely to be of some significance. In the present case comparisons have not been possible on a very wide scale although in the evaluation of fragment performances it has been possible to make some very useful comparisons across the three main data samples under consideration i.e. the 20 amino acids, the 39 local <sup>a</sup>aromatics and the 79 penicillin structures.

#### 3.4.1. Method

##### 3.4.1(a) The Association Measure

The classifications were carried out using Dice's SC and

the binary representation of structures based on additive coding described in the previous section. In this earlier study simple matching coefficients were shown to perform as well as, and in some cases better than the more detailed association measures based on quantitative fragment descriptions. One of these measures was therefore applied in the present study as they require shorter numerical representations than the quantitative measures, and this was an important consideration here, because of the larger numbers of descriptors arising with larger fragments and fragment combinations.

#### 3.4.1(b) The Structure Representation

Details of the numerical representation used and the way in which this is derived from a redundant connection table record are given in the previous section.

#### 3.4.1(c) The Fragments

A wide range of atom and bond-centred fragments may be extracted from the connection table record, and the particular definitions examined here are some of the fragment types which have already been investigated elsewhere as screens for substructure searching.<sup>208,212,213</sup> Four bond-centred and two atom-centred fragments were considered in all, and in each type the fragments describe progressively larger regions of the molecule around the central bond or atom. The two atom-centred fragments used were simple atom descriptions and augmented atoms, the second of which was used for the investigation of association measures and numerical representations described previously. The four bond-centred fragments, referred to in previous publications<sup>212,213</sup> as 'simple pairs', 'augmented pairs', 'bonded pairs' and 'octuplets', show a more gradual expansion from the central bond. Thus, simple pairs describe the central bond and the atoms it connects, augmented pairs

describe, in addition the terminal connectivities of the atom pair, bonded pairs describe the external bond types, and finally octuplets describe both the external bonds and the external atoms i.e. the atoms connected to the central atom pair. Figure 19 gives the different substructures arising in three of the penicillin side chain structures. As these examples show the three largest bond-centred fragments have the advantage that they can identify ortho ring substituents and separate these from meta and para ring derivatives. Larger fragments would be required to distinguish meta and para disubstituted structures, and other more detailed substitution patterns.

### 3.4.2 Fragment Performances

#### 3.4.2(a) Prediction Levels

Property predictions were simulated in the manner described in previous sections, and summaries of the deviations between observed and predicted property values obtained in each sample are given in Tables 5, 6 and 7. Sum of squares ratios were calculated as before, and in Figures 19 to 24 these are plotted against sample to feature ratios to illustrate the different levels of prediction obtained by each fragment.

##### (1) Nearest Neighbours

The predictions based on highest SC values in each sample are given in Figures 19, 21 and 23. As shown, the greatest variation in fragment performances arises in the structurally diverse group of local anaesthetics, and the least variation in the penicillins.

Some interesting differences are shown in the two smaller samples. In the amino acids augmented atoms perform very well,



and the larger bond-centred fragments give slightly better performances than the smaller ones. Simple pairs and bonded pairs combined show a slight improvement over the separate performances of these fragments, and octuplets combined with these show a further improvement, although the result in this second case is not as good as the prediction given by octuplets on their own. In the anaesthetics opposite trends are shown, with atoms and the smaller bond-centred fragments giving better performances than the larger fragments. Atoms give the lowest sum of squares ratio, and simple pairs show a marked improvement over bonded pairs and octuplets. The combined fragment performances in this case lie close to the average performances of the different fragments involved.

In contrast with these two groups, the different fragment performances in the penicillins lie within a very close range, with simple pairs giving the lowest sum of squares ratio of 0.34 and octuplets the highest value of 0.51. The combined fragment performances are again close to the average performances of the individual fragments involved.

(ii) Classifications

Predictions based on the classifications are shown in Figures 20, 22 and 24. These are very similar to the predictions given by nearest neighbours, although a few differences arise, and the most notable of these occur in the amino acids. In this group the larger fragments again perform best, but there is now a more marked separation between octuplets and the remaining bond-centred fragments, with this fragment giving an improved result and the others much lower levels of prediction, and variance ratios in the

region of unity. As before, simple pairs and bonded pairs combined lead to a slight improvement, but when octuplet descriptions are included with these a very poor prediction results, despite the improved result given by octuplets alone. These differences are discussed below.

In the <sup>a</sup>anesthetics, there is, as with the nearest neighbours, a gradual improvement in the prediction level as the fragment size decreases, with atoms giving the lowest sum of squares ratio of 0.21 and octuplets the highest value of 0.67. The range is slightly smaller in this case.

A similar, very small range is covered in the penicillins, with augmented atoms giving the lowest sum of squares ratio of 0.49, and simple pairs and bonded pairs the highest value of 0.65. The levels of prediction are slightly lower in this case.

The combined fragment performances in the two larger samples are very similar to the nearest neighbour results.

From a prediction point of view therefore the fragments behave differently in each sample. The amino acids and local anaesthetics show opposite trends in fragment performances, and the penicillins, which have some of the characteristics of the closely related group and some of the structurally diverse group, show a fairly constant level of prediction. However, many of the differences arising in the two smaller samples are quite small, and with a few exceptions they are unlikely to be of statistical significance. In addition, the reliability of the predictions for comparative studies, as discussed in Section 3.3, will depend on the particular relationships arising in each sample between structure and property data. In view of these difficulties, and

the absence of rigorous evaluation procedures, it is therefore important that the differences between samples be judged in relation to the types of structural arrangements produced. These are discussed below.

#### 3.4.2(b) Structural Arrangements

In each sample all of the fragments produce sensible chemical arrangements, with the exception of some of the clusters obtained using atom descriptions. A few interesting variations arise but in each sample essentially the same arrangements are produced and examples of these are given in Figures 25 to 32.

In the amino acids some of the more closely related structures, such as the two acidic amino acids, aspartic and glutamic acid, and the amides, asparagine and glutamine, form well defined clusters in every case. The cyclic and acyclic derivatives are also clearly separated in most cases, although the non-aromatic ring derivatives are not always associated with the cyclic structures, and chain onto the acyclic group. The aromatic derivatives are well defined, even when simple atom descriptions are used. Using bonded pairs, however tryptophan does not join this group, due to the more important influence of the partially saturated 5-membered ring system in this case, and it clusters instead with histidine, which incorporates a similar 5-membered ring system.

Other similar structural types arising in the acyclic group, which are not quite as closely related as the structure pairs mentioned above, are not clearly identified in every case, and are only clustered when more detailed fragment descriptions are used.

For example, the basic amino acids, lysine and arginine cluster when octuplets and augmented atoms are used, but are not strongly associated in other cases. Using simple pair, augmented pair and bonded pair descriptions these structures chain at very low levels of association and the very poor predictions obtained for them account for the low prediction levels in these classification cases. They also account for the poor prediction levels in the combined fragment cases, and explain why octuplets combined with simple pairs and bonded pairs give a much less satisfactory prediction level than octuplet descriptions alone. Atoms give a slightly better prediction because lysine and arginine now cluster with the cyclic derivative, histidine, which has similar basic properties. This is an example of one of the more satisfactory associations arising from atom descriptions, as although these fragments have failed to identify the gross structural features present, they have successfully identified the important NH groups in these structures which account for their high basicity.

Other similar acyclic structures which do not form a definite cluster in each case are the long chain alkyl derivatives, leucine, isoleucine and valine. Again the larger fragments, augmented atoms and octuplets, have been successful in identifying these, but using augmented pairs and bonded pairs the structures are not strongly linked. Leucine and isoleucine come together again when atoms and simple pair descriptions are used, due to the inability of these fragments to distinguish between them, and augmented atoms, as discussed previously, are also unable to separate these two isomers.

Using atom descriptions a number of unsatisfactory associations arise between cyclic and acyclic derivatives of comparable size, although, as seen from the example quoted earlier, these associations are not necessarily unsatisfactory from a prediction point of view. Another example of this is the cluster formed between proline and valine, both of which are only slightly acidic. Atom descriptions are unable to separate these two structures, and like leucine and isoleucine these cluster at SC level 1.

The combined fragment descriptions give results very close to the performances of the original fragment descriptions and they show that the additional information in these cases has not led to any serious distortion in the relationships between structures. Simple pairs, bonded pairs and octuplets combined give clusters which are almost identical to those produced by octuplets alone, and the reason for the wide discrepancy between the prediction levels in these two cases is discussed at the beginning of this sub-section.

In the local anaesthetics, atoms give rise to a much larger number of associations between quite unrelated structural types, but the other classifications give more satisfactory clusters from a structural point of view. The classification produced by augmented atoms has been described in detail in Section 3.3 (Figure 11), and the bond-centred fragments give results which are very close to this. They all show an early breakdown into cyclic and acyclic classes. The smaller acyclic group shows a well defined cluster of normal alcohols, but depending on the fragment size, methanol and isopropanol are not always closely associated with this group. Thus methanol has no augmented pair fragments, bonded

pairs or octuplets in common with the larger alcohols, and is dissociated from the group in these cases. Using simple pairs, on the other hand, it now has one feature in common with the rest of the group and joins it at a lower level. Simple pair descriptions are also unable to distinguish between n-propanol and isopropanol, and these cluster at SC level 1, before joining the main alcohol cluster at a slightly lower level. Using the larger bond-centred fragments, isopropanol, like methanol, is dissociated from the main cluster. When simple pairs and bonded pairs are combined methanol again joins the alcohol group, because of the simple pair feature it has in common with the group. However, because of the larger numbers of descriptors involved in this case its association with ethanol is now very much weaker than in the simple pair case.

The larger cyclic group also shows some clearly defined and chemically sensible clusters in each case, which correspond closely to those produced by the simple matching coefficients discussed in the previous section. Thus, simple benzene derivatives<sup>v</sup> form a separate cluster, or are reasonably closely associated, as are the fused ring derivatives quinoline, 8-hydroxyquinoline and O-phenanthroline. In each case, as before, benzimidazole<sup>z</sup> does not associate with this second group because of the influence of the 5-membered ring, which is coded as a localised ring system, and the remaining heterocyclic compounds are also dissociated from this group because of the larger and more important influence of acyclic components in these cases. In addition, the cyclic group shows some well defined clusters between the structures involving this type of component, i.e. the previously discussed dialkylamino derivatives of acetanilide and the dialkylamino ethyl ester derivatives of benzoic acid. In all

cases the two structural isomers, phenyltoloxamine and diphenhydramine, are not strongly linked with this group for reasons given in Section 3.3, and in each case these two isomers form a separate cluster at a high level.

The arrangements of structures in the cluster of cyclic compounds is roughly the same in each case. Simple pairs give a slightly different breakdown initially, but these also give the same basic clusters at higher levels. Using the larger bond centred fragments, and fragment combinations the breakdown corresponds closely to the structuring produced by augmented atoms i.e. there is an initial breakdown between rings with large and small acyclic components, with the first group dividing according to the nature of the chain structure, and the second according to the type of ring system. Using simple pairs the initial splitting depends more on the type of ring system present, which means that the simple benzene derivatives are now more closely related to structures such as xylocaine and procaine, than to the fused ring derivatives involving small ring substituents, such as quinoline, 0-phenanthroline and 8-hydroxyquinoline. However, the type of chain component has also had some influence in this case, bringing structures such as dibucaine, caramiphen and eserine clearly within the bounds of the simple ring group, due to associations with xylocaine, tetracaine and procaine. Diphenhydramine and phenyltoloxamine are also more closely related to this group than to the fused ring group.

One noticeable effect on the classification as the fragment size increases is the increased degree of chaining, particularly with some of the smaller acyclic structures which share very few

features with the rest of the acyclic group. As the fragment size increases, progressively larger numbers of these chain at SC level 0, or at a very low level of association, and because these structures are very poorly predicted this accounts for the gradual lowering in the prediction level in moving from simple pairs through to octuplets. For example, structures such as methanol, acetone and propanol form strong associations in the acyclic group in the simple pair case and are all reasonably well predicted in this case. Isopropanol which clusters separately with propanol is also well predicted. Augmented pairs give a slightly less satisfactory result for these structures, and methanol is now completely dissociated from the acyclic group and is very poorly predicted. Acetone and isopropanol are also dissociated, but form a separate cluster further along the hierarchy and are well predicted as a result. Propanol remains within the acyclic group but now associates with the higher alcohols present, all of which have much lower log (MBC) values. Finally, using the two largest bond-centred fragments acetone, methanol and isopropanol all chain at very low levels and are all very poorly predicted. An interesting result with octuplets compared with the smaller fragments is that urethane and ethyl ether, which leave the acyclic group when augmented pair and bonded pair descriptions are used rejoin it again in this case. This leads to a slight improvement in the prediction levels, but these are not as good as the predictions obtained for these structures using simple pairs. This is because of the absence in the acyclic group at the octuplet level of description of the other smaller acyclics of lower activity, and the resulting much lower group average log (MBC) value, on which the prediction of these



structures is based. Examples of some of the above predictions are given in Table 8.

In contrast with the above fragment types atoms do not give a clear separation of cyclic and acyclic derivatives and fail to identify the groups which are usually regarded to be of chemical interest, such as the normal alcohols, and the simple benzene derivatives. They are unable to separate the various structural isomers present, such as propanol and isopropanol, and phenyltoloxamine and diphenhydramine, and they also give some unsatisfactory associations between ring and chain structures of comparable size e.g. hexanol and phenol, and heptanol and benzyl alcohol. Other structures they are unable to separate are those which have identical molecular formulae except for the number of hydrogen atoms present, for example, thymol and 2-naphthol.

Many of the associations, however are good from a prediction point of view. This is because the hydrophobic and hydrophilic groups important for activity are present in a wide range of structural types, and atoms have been able to recognise many of the structures with similar functional groups, without identifying the wider structural differences which would normally separate these. For example, eserine which has no strong associations in other cases, and is poorly predicted by the larger fragments, clusters with tetracaine, which has a dissimilar ring system, but a reasonably close level of activity. These structures have a number of important features in common, such as similar amine, tertiary amine, carboxy and aromatic groups. The structures which lie closest to eserine on the activity scale in question are quinine and caramiphen, but these are not as strongly associated due to larger numbers of carbon atoms. Quinine is also well predicted.

With the more detailed fragment descriptions this saturated bridged ring structure chains at a low level, but using atoms it clusters with dibucaine, which has a similar high level of activity. The two structures have dissimilar characteristics overall, but they share a quinoline ring system, and they both show the features considered important for local anesthetic activity i.e. the presence of an aromatic component separated from a hydrophilic group by a carbon chain.

Some of the smaller, less active structures have also been better predicted in this case, such as thymol, 2-naphthol and diethyl ether. The two phenols form a separate cluster and are well predicted. These chained in the case of the larger fragments. Diethyl ether clusters with butanol, which has a very close level of activity, and this previously formed only very weak associations in the acyclic group. Butanol, which normally clusters with the higher alcohols, is also better predicted in this case, and hexanol, which clusters with phenol is another of the alcohols which is well predicted. Not all of the associations are more successful for prediction, for example, benzyl alcohols association with heptanol instead of the simple benzene derivatives, results in a very poor prediction for this structure. However, the large number of good associations has resulted in an overall improvement in the level of prediction, and examples of these are given in Table 9.

In the penicillin sample, the penicillin nucleus has had very little influence on the classifications produced except to increase the overall levels of similarity obtained between structures. The clusters have been determined largely by the nature of side chain structures, and as these cover quite a wide

range of structural types the sample in this respect resembles the structurally diverse group of local anaesthetics. Another point of similarity between the two samples is that similar levels of activity are often shown by quite dissimilar chemical types.

The penicillin side chain structures may be divided into five main categories, namely, a small group of acyclic structures and a larger group of cyclic compounds which divides into groups of simple benzenes, naphthalenes, quinolines and thiophens. Each of the fragments used, except atoms, have been successful in identifying these different groups, and similar overall arrangements have been produced in each case. Octuplets fragments give a slightly sharper resolution of clusters, and the naphthalene structures in this case are broken down into a number of smaller groups which are separated to some extent. The remaining groups are still clearly defined, however, and the only other difference in this case is the separation of the thiophen group from the remaining cyclic structures by the small acyclic group.

Within each of the main groups the fragments show a few variations which could be of importance, the main differences being found in the simple benzenes, which make up the largest part of the cyclic cluster. A large proportion of the structures in this group are simple halogen derivates, and the smallest fragment used i.e. simple pairs, tends to cluster these according to the number of halogen substituents present. Thus, the non-substituted derivates are separated from the monohalogen derivates and these in turn are separated from the di- and tri-halogen derivatives. The different halogens are separated, but no distinctions are drawn between ortho, meta and para ring deriva-

tives of a similar type, and these often cluster together at SC level 1, e.g. the ortho, meta and para fluoro substituted phenoxymethyl derivatives 52, 53 and 54, and the ortho, meta and para fluoro substituted  $\alpha$ -phenoxyethyl derivatives 60, 61 and 62. At the other end of the scale octuplets tend to cluster the simple benzene group according to the type of benzene derivative in question, and in particular the nature of the connecting side chain to the parent structure. This latter feature is often of predominant importance, and brings structures involving different numbers of halogen substituents within the same cluster. For example, a well defined cluster is formed between the various chloro substituted  $\alpha$ -methoxybenzyl derivatives, and another cluster is formed between two  $\alpha$ -aminobenzyl derivatives, one of which involves a single chlorine substituent in the para position to the  $\text{CH}_2$  group and the other, two chlorine substituents in the meta and para positions. The various mono- and di-halogen derivatives which involve similar connecting side chains are again clustered together as in the simple pair case, but now a distinction is drawn between the ortho and the meta and para ring derivatives. The latter pair are still inseparable and the ortho derivative usually joins these at a slightly lower level.

The remaining, medium-sized fragments perform somewhere in between these two extremes. The augmented atom classification corresponds more closely to the simple pair result, and usually brings together the structures involving similar numbers of halogen substituents. It is also unable to distinguish between ortho, and meta and para ring derivatives. In this case however, sharper distinctions are made between the different types of

benzene derivatives, for example, simple pairs are unable to separate the structural isomers 41, 42, 57, 58 and 59, some of which are dichloro substituted  $\alpha$ -methoxybenzyl derivatives, and others dichloro substituted  $\alpha$ -phenoxyethyl derivatives, but these two groups are clearly separated in the augmented atom case.

The bonded pair result is similar in many respects to the octuplet result, but it also retains some of the characteristics of the augmented atom classification. Thus, many of the clusters formed involve only mono- or di-halogen derivatives but in this case, the position of the halogen group has often been a more significant factor than the type of group. This has given rise to a number of mixed halogen groups, such as the cluster formed between the  $\alpha$ -methoxybenzyl side chain structures 39, 40 and 45, which are meta substituted bromo, chloro and fluoro derivatives respectively. Another example is the cluster formed between the two  $\alpha$ -methoxybenzyl derivatives 42 and 44, which are both substituted in the meta and para positions to the  $\text{CH}_2$  group, and one of which is a dichloro derivative and the other a chlorofluoro derivative. Halogen derivatives of a similar type, however, are still brought together in cases where the overall structural features are similar, although as with augmented pairs and octuplets ortho substituted derivatives are now distinguished from meta and para derivatives. In this case quite wide separations have often arisen between these isomers. For example, the ortho fluoro substituted  $\alpha$ -phenoxyethyl derivative no longer clusters with the meta and para derivatives in this group but clusters instead with the ortho chloro substituted derivative. Another example is the ortho fluoro substituted phenoxyethyl derivative, which is more

closely associated with the non-substituted derivative in this group than with the corresponding meta and para derivatives.

In the simple benzene cluster therefore there are a number of basic similarities between each of the classifications, but there are also a number of important differences. Each of the results, is sensible on a chemical basis, and the variations arising have had very little effect on the overall levels of prediction obtained. From a structure-property viewpoint therefore it is impossible to say which of the fragments, if any, is of greatest value. It is also difficult to say which of the structural arrangements is the most satisfactory. However, the differences here could be quite important in a retrieval situation, and the choice of suitable fragment in this case could depend on the particular application.

Atom descriptions produce a very different classification result, and as in the anaesthetics they give rise to a larger number of associations between quite dissimilar structural types. Some of the very close associations have been retained, for example, between the various structural isomers, such as the simple benzene derivatives 23 and 24, and 32 and 33, and the naphthalene derivatives 68, 69 and 70. A number of the closely related halogen derivatives have also been clustered, as have several of the acyclic derivatives. However there is no longer a clear division between the different chemical groups discussed above, and a large number of associations arise between dissimilar ring types, and between ring and chain structures of a similar size. For example, the n-heptyl side chain derivative now clusters with the non-substituted benzyl derivative, and the

$\alpha$ -amino derivatives of these two side chain structures also cluster together. The longer alkyl chains cluster with the larger ring derivatives, and some of the smaller ring systems incorporating alkyl side chains also make associations with these, such as the association between the  $\alpha\alpha$ -diethylbenzyl derivative and one of the naphthalene derivatives.

The associations using atoms have led to better predictions in a few cases, but there have been fewer improvements in this case compared with the anaesthetics, and there has not been an overall reduction in the level of prediction obtained. The less satisfactory associations are due to the fact that the structures which are most similar on a size basis, in this particular sample do not incorporate the most similar functional groups as often as they did in the anaesthetics. However, the associations have not led to a noticeable drop in the prediction level, and it is much more difficult in this sample to rationalise the structure-activity relationships, and to explain the fairly constant level of prediction in terms of individual associations and the hydrophobic and hydrophilic groups arising, which are thought to be important for serum binding.

### 3.5 Conclusions

The results of these investigations show that the combination of structure handling techniques originally developed for information storage and retrieval, and numerical taxonomic techniques developed for biological classification, lead to classifications which are sensible from a general qualitative chemical point of view. The methods developed give substantial agreement between the classifications and SCs and DCs

based on the structure diagram, and the available physical and biological properties. Predictions simulated on the basis of the classifications, SCs and DCs were found to be in good agreement with observed property values. This result is encouraging as it shows that structure-property relationships implicit in the data can be brought out without the construction of a physical model defining explicitly the relationships between the structure diagrams and the observed properties. The results therefore suggest that the method could be valuable for property prediction, as well as for file handling purposes.

The initial investigations showed that it is possible to obtain sensible clusters under very simple conditions, applying simple matching coefficients to binary representations of the structure diagram. They indicated however that this type of numerical representation of structures is feasible provided the different occurrences of different fragment types in a structure are described and a distinction is made between cyclic and acyclic substructures. These details were therefore included in all subsequent investigations.

In the comparison of resemblance measures the results obtained for 39 local anaesthetics showed that it is possible with a relatively simple classification approach to obtain meaningful chemical groupings in a quite diverse range of structural types. Most local anaesthetic agents considered acceptable for clinical purposes incorporate a lipophilic aromatic residue and a hydrophilic amino group connected by an intermediate hydrocarbon chain. It has been shown, however, that a wide variety of structures exhibit local anaesthetic activity and in the present sample structures range from very



simple aliphatic molecules such as methanol and chloroform to fairly complex structures such as eserine and caramiphen.

The performances of the similarity and dissimilarity coefficients and the classifications derived from them were compared, as in the investigation of numerical representations, using the relationship between structure and property to simulate the prediction of log (MBC) values. However, the difficulty with this approach in structurally diverse samples in which similar chemical types do not always have similar activities, is that prediction levels do not necessarily reflect the type of classification arising. Therefore this may not be the most useful method of evaluation. Even so, some agreement was obtained in the present sample between predictive performances of coefficients, and the classifications derived from them, and the type of structural arrangements produced. The best results were obtained for the simple matching coefficients, based on binary representations of fragments, and the simple distance function which uses a quantitative fragment description. These measures gave better classifications and also better predictions than the functions involving probabilistic weighting, where the significance of each fragment is related to its probability of occurrence. The simpler measures also showed that the coefficients based on more detailed quantitative fragment descriptions performed no better than those based on simple binary representations in which fragment occurrences are taken into account. Another important observation was the very poor result given by the standardised distance function compared with the non-standardised measure, and this suggested that

the standardisation process had masked the important differences between fragments. Thus, both character standardisation and character weighting procedures used in this study have an adverse effect on the result.

The amino acids and penicillins showed similar relative performances between coefficients, and this suggests that the differences indicated between measures are of some significance. However, in each sample, many of the predictions were close, and considering the small scale of application and the very approximate method of evaluation used, additional studies would be required to verify the observed trends. The probabilistic measures performed particularly poorly, and in this case it would be useful to investigate whether other weighting criteria would be more appropriate, for example information-theoretic character evaluations, where the significance of a character is related to the total number of alternatives possible for that character, expressed in terms of a probability function.

In the final investigation, a simple matching coefficient and binary representation were used to test the classification performances of a number of two-dimensional substructural definitions. Again, sensible chemical arrangements were obtained using a wide range of definitions, and very simple fragment descriptions were found to be as effective as more detailed definitions in some cases. On this occasion, however, there was not always good agreement between the chemical significance of classifications and their predictive performance. For example, atom descriptions gave the least satisfactory structural arrangements

and poor predictions in the amino acid and penicillin samples, but a good prediction in the structurally diverse group of local anaesthetics. In this sample the functional groups which are important for activity often arise in compounds showing quite wide structural differences overall, and atoms have been more successful in some cases in identifying these groups. However, as the results obtained for the two larger samples show, this is not necessarily the case and the suitability of this fragment will depend on the particular relationships arising between structures in the data sample in question.

The remaining fragments gave similar classifications in each sample but/again showed some interesting variations in predictive performance. The wider differences arising in the two smaller samples could be explained satisfactorily. These were not the outcome of wide differences in structural arrangement, and the prediction levels in these groups appeared to be a lot more sensitive to small changes in the classification than the levels produced in the larger penicillin sample. This suggested that the differences were not of any real significance. On the other hand there appeared to be some consistency in the two smaller groups between prediction and classification performances, with the fragments giving the best predictions also producing a slightly sharper resolution of structures, and this could mean that the differences shown between samples, were of some importance. However, as each sample has performed differently and there are no statistical tests available to compare them, here, as in the comparison of coefficients, investigations would be required on a much wider scale, using a much larger range of structures and properties, before any

trends in the data could be treated with confidence. Without more definite guidelines of this nature, the present investigation illustrates the importance of considering a range of descriptors for any particular application. As in the previous study of coefficients, the classifications obtained by many of the fragments were close, and the 'best' of these from a chemical point of view could depend on the application or even the interests of the investigator. Finally, it has been shown that the predictive value of the classification may not be the most useful criterion for evaluating method performance, and, depending on the aims of the classification, the structural arrangements could be more useful, especially in comparative studies. However, the results suggest the relationships between structure and property data using small samples may be influenced by the particular structures and property-values arising, and in larger samples more consistent results and a closer agreement between the predictive performances of clusters and their chemical significance may be obtained.

CHAPTER 4

The Development of an Empirical Structure - Property

Correlation Method based on Regression Analysis

This chapter describes the work carried out on a new empirical structure-property correlation method based on regression analysis. Property data is related directly to the structural features of the molecule, and this is the first statistical model of this type reported to employ automatic procedures of substructural analysis, and to relate property data to the structural characteristics of the complete molecule.

The approach is simple, and its advantage over other parametric approaches based on regression analysis have been discussed in Chapter 2. Probably the most important feature of the method is its utilization of all the structural features present, and its resulting ability to handle structures which do not belong to the same chemical series. As discussed earlier, most other approaches are restricted to the investigation of side chain structures and to the problem of property optimisation within a given lead series. In a recent investigation Nys and Rekker<sup>214</sup> have also looked at a wider range of structural features, using a similar regression model to determine fragment  $\Pi$  values. However, they do not base their analysis on an automatic breakdown of the structure diagram, and structural features are not investigated systematically, but chosen on the basis of assumed chemical significance. One of the problems with the more usually considered semi-empirical structure-property methods which rely on physicochemical property data is the reliability of calculated property values and the frequent need to measure these directly. This is no problem in the empirical case, and with the increasing costs of preparing new compounds, especially in the pharmaceutical industries,<sup>158</sup>

this may be an important factor governing the choice of methods in future applications.

The model developed is subject to the usual limitations associated with the use of regression analysis, and the important restriction on the sample to feature ratio has meant that investigations have not been possible on the small group of amino acids. In this sample the variety of substructures exceeds the number of structures even when very small fragments are considered. The local anaesthetics and penicillins are more suitable data sets, and it has been possible in these to consider a variety of substructural definitions. As the number of substructures increases with the fragment size, a wider range of fragments could be considered in the larger of the two groups. In both samples it has also been possible to extend the range slightly in cases where the number of substructures does not greatly exceed the number of structures, by discarding features common to each structure, and features which always arise together. This has no effect on the regression solutions. Where necessary some very highly correlated variables have also been excluded and this should affect the result only slightly, because of the large numbers of variables usually employed.

An attempt is made to estimate the power of the method for property prediction by simulating the prediction of unknown property values using the 'hold-one-out' technique. This, as discussed in Chapter 2, gives a more realistic assessment of predictive power than the regression equations in which all the structures are included, as each structure in the second case is influenced by its own observed

value. In applying this technique the structures under investigation are partitioned into two sets, the 'test set' consisting of the single structure to be predicted and the 'design set' consisting of all other structures in the group. The regression analysis is then carried out on the design set and the regression constant and coefficients obtained are used to 'predict' the property of the structure in the test set.

In the earlier attempts to correlate structure and property data mathematically, simple linear combinations of structural or physicochemical parameters were usually considered. However, it eventually became evident that the addition of interaction terms to such equations could sometimes lead to a better correlation.<sup>215-217</sup> In the present case higher order terms would take into account very approximately the interaction between substructures and it is possible that these will lead to similar benefits in the correlation. Two such expressions have therefore been considered to see whether these give a significant improvement over the linear result.

As with the classification work, the results obtained from such small investigations as these may not be of general significance. However, in this case it is possible to say whether the results are significant from a statistical point of view, and this puts the approach at a considerable advantage over the non-parametric methods. An indication is also given of the likely contributions to activity of individual substructures and a very important feature of the method is that it is possible to estimate the significance of these contributions and to compare them on a statistical basis.



4.2 The Empirical Model

In the basic regression model it is assumed that the property under investigation,  $y$ , of the  $i$ th compound is related to the structural features present in this compound by the expression,

$$(1) \quad y_i = \sum_{j=1}^n b_j x_{ij} + \text{const}$$

where there are  $n$  types of structural fragments in the set of structures, and  $x_{ij}$  is the number of times that the  $j$ th fragment occurs in the  $i$ th structure. The regression coefficient,  $b_j$ , for the  $j$ th feature represents the contribution of this feature to the property in question, and both this and the regression constant (const) are determined by the analysis.

In the anaesthetics the following two expressions were also considered

$$(2) \quad \log(\text{MBC})_i = \sum_{j=1}^n | b_j x_{ij} + c_j (x_{ij})^2 | + \text{const}$$

where  $b_j$ ,  $x_{ij}$  and const are as defined above in (1), and  $c_j$  is the regression coefficient of the squared term for the  $j$ th fragment.

$$(3) \quad \log(\text{MBC})_i = \sum_{j=1}^n b_j x_{ij} + \sum_{j=1}^{n-1} \sum_{k=j+1}^n d_{jk} (x_{ij} x_{ik}) + \sum_{j=1}^n c_j (x_{ij})^2 + \text{const}$$

where  $x_{ij}$ ,  $b_j$ ,  $c_j$  and const are as defined above,  $x_{ik}$  is the number of times of the  $k_{\text{th}}$  fragment occurs in the  $i_{\text{th}}$  structure, where  $1 < k < n$ , and  $d_{jk}$  is the coefficient for the cross product term relating to fragments  $j$  and  $k$ . The first of these expressions is a quadratic which includes only squared terms, whilst the second includes both these terms and cross-product terms, i.e. a full quadratic.

#### 4.3 Method

##### 4.3.1 The Input Matrix

The regression analysis was carried out automatically using a computer manufacturer's statistical analysis programs,<sup>219</sup> and an example of the type of input matrix required by the regression package is given in Figure 33. This is also derived automatically. In it structures are identified by row numbers, and for each structure a frequency vector is set up giving the frequency of occurrence of each substructure present in that structure. These vectors are very similar to the 'distance' vectors, used for the computation of distance coefficients in the classification work, and they are obtained by a very similar process. Thus, structures are input to the computer as redundant connection table records and these are first analysed to determine the different fragment types occurring. The fragments are then listed and used to set up the required frequency vectors for regression. Property values are included in the matrix as an additional variable, and the information to be used as the dependent variable, in this case the property values, are identified on input to the regression programs.

##### 4.3.2 Dependent Variables

The property data used as the dependent variable in these investigations were the local anaesthetic and serum binding property values used previously to test the predictive performances of the classifications. As before, serum binding values were considered in the form  $\log(B/F)$ , where the amount of penicillin bound to human serum (B), is taken as a ratio of the amount left free, (F). Thus structural features with positive regression coefficients increase serum binding, whereas they decrease local anaesthetic activity.

#### 4.3.3 Independent Variables

The different substructures used in this investigation have already been described in Chapter 3.

In the anaesthetics, regressions were carried out using atom, simple pair and augmented pair descriptions. Attempts to use larger fragments in this smaller sample failed because the number of variables required to represent the structures in these cases greatly exceeded the number of structures. An automatic analysis of the 39 structures showed them to contain 4 different atom fragments, 16 simple pairs and 43 augmented pairs. The augmented pair fragments were reduced to the required limit by excluding a number of perfectly correlated fragments. Two groups of three fragments and three groups of two fragments were found to have within group correlation coefficients equal to one, and by excluding all but one of the fragments from each of these groups the number of variables was reduced to 36.

In the penicillins, augmented atoms, bonded pairs and octuplets were considered in addition to the fragments used above. Excluding where they arose the fragments of the parent structure which occurred with the same frequency in every structure, the group was found to contain 7 different atom fragments, 14 simple pairs, 33 augmented pairs, 45 augmented atoms, 51 bonded pairs and 97 octuplets. The octuplet set was reduced to the necessary limit by excluding perfectly correlated fragments, as described for the anaesthetics. It was possible to exclude 26 such fragments in all, giving a final total of 71 variables.

The quadratic expressions used in the anaesthetics involve much larger numbers of variables and it would not have been possible to consider these in every sample and for every substructure. In this sample they gave rise to too many augmented pair variables for analysis, leaving atoms and simple pair descriptions. Atoms, however, do not show up any structural characteristics and the expressions were considered to be of less value with this definition. They were therefore only applied in the simple pair case.

Using expression (2) it is only necessary to introduce squared terms for the variables for which more than two different values occur in the given set of structures. Ten of the simple pair fragments arising in the anaesthetics fall into this category and including these gave a total of 26 variables in all. Expression (3) gives rise to a total of 152 variables, and of these 69 occur with the same frequency in every structure and could be considered as constants. Another 39 variables which belonged to groups of perfectly correlated variables were also excluded. This reduced the number of variables to 44, which was finally reduced below the required limit by excluding another 8 variables belonging to groups of highly correlated features, where the intra-group correlations exceeded 0.9. Again in this case one variable from each group was retained.

#### 4.4 The Correlations

##### 4.4.1 Structure-Property Agreement

The regression analyses were carried out at two different levels of significance i.e. by including in the regression set variables which will give a fit at two different levels of confidence. At the first level, usually referred to as the 99% level, there is effectively no

confidence limit, and each independent variable is forced into the regression set, providing its pivot element satisfies certain basic criteria.<sup>218</sup> At the second level a stepwise procedure is followed, in which structural fragments are introduced in decreasing order of their pivot elements, and only those with coefficients significantly different from zero at the 10% level are retained.

The different correlations arising in each sample are compared, and the significance of each correlation determined using the F-test. To test the significance of individual correlations F-values were computed as follows:

$$F = \frac{R^2 (n-m)}{(1-R^2)(m-1)}$$

where R is the multiple correlation coefficient, n is the numbers of structures, (m-1) is the number of independent variables included in the regression ( $\gamma_1$ ), and (n-m) is the numbers of degrees of freedom ( $\gamma_2$ ). For two correlations obtained within the same data sample, F values were computed as follows:

$$F = \frac{(re_2)^2}{(re_1)^2} \quad (\gamma_2, \gamma_1)$$

where  $re_1$  is the residual error obtained using correlation 1,  $\gamma_1$  is the number of degrees of freedom for correlation 1 and  $re_1 < re_2$ .

From the values of F,  $\gamma_1$  and  $\gamma_2$  the corresponding significance level can be found from statistical tables by checking against the appropriate threshold values of the F distribution. The 5% significance level is usually considered the lowest limit of confidence for general statistical purposes, and correlations which did not differ significantly at this level were considered identical. At this level there is a 1 in 20 probability that the given result could have arisen by chance.

Summaries of the analyses obtained in each sample are given in Tables 10 and 11. In all cases high correlation coefficients were obtained and F values were of high significance. The results at the 10% level do not differ significantly from those at the 99% level, although these give a slightly lower residual error in most cases and therefore provide a slightly better explanation of the data. The correlations at this level are discussed below.

In each sample there is a gradual improvement in the result as the fragment size increases. Thus the lowest residual error in the anaesthetics is obtained using augmented pair descriptions, and the agreement between structures and properties in this case is slightly better than in the simple pair case. The two correlations differ at the 5% level. Atom descriptions give a very poor residual error in comparison with these two fragments, and the correlation in this case differs significantly from the simple pair result at the 1% level, and from the augmented pair result at the 0.1% level.

Quadratic terms introduced with simple pairs do not lead to a noticeable improvement. Neither of the correlations differs significantly from the linear result at the 5% level, and when only squared terms are introduced (expression 2) the residual error is increased slightly. The full quadratic (expression 3) gives a better result and leads to a marginal improvement over expression 1. However, both correlations again differ from the augmented pair result at the 5% level, and in the case of expression (2) a difference is indicated at the 1% level.

The penicillins show a similar gradual improvement in the correlation as larger fragments are considered. A few differences arise

between the medium sized fragments and more significant differences in the extreme cases. Atoms again give the least satisfactory correlation and the highest residual error. Simple pairs lead to only a slight improvement in this case, and the differences between these two correlations is not significant. Augmented atoms, augmented pairs and bonded pairs all give lower residual errors, and correlations which differ from the atom and simple pair results at the 1% or 0.1% levels. These medium sized fragments give very similar results and no differences are indicated between them at the 5% level. Octuplets give the lowest residual error and thus the most satisfactory explanation of the data. The correlation in this case differs from the atom and simple pair results at the 0.1% level, and the augmented atom and bonded pair results at the 5% level. However, augmented pairs, which give a slightly lower residual error than bonded pairs and augmented atoms do not differ significantly from the octuplet result.

#### 4.4.2 Use of the Correlations for Property Prediction

Property predictions, simulated by the 'hold-one-out' technique, were carried out for each of the local anaesthetics, and for a random sample of 20 of the penicillins (see Appendix 1). Augmented pair descriptions were used in the local anaesthetics. In the penicillins, in addition to considering the best correlation result using octuplets, predictions were also carried out with some of the medium sized and smaller fragments to see in which way the different agreements between structures and properties influence the levels of prediction obtained.

A property value was obtained for each test set structure by summing the appropriate regression coefficients from the design set, as shown in the examples given in Figures 34 and 35. Fragments present in the test set structure which are absent in the design set, or which have been excluded from the regression, are assumed to have zero coefficient values.

The predictions for the anaesthetics are summarized in Table 12. This shows that the regression coefficients obtained at the 10% significance level give a much more satisfactory result, and the predicted values at this level are listed in Table 13. Eight of the structures present contain unique augmented pair fragments, which meant that insufficient parameters were available for prediction from the analyses which excluded them. In these cases it is necessary either to estimate the missing values or to assume they are zero. They were assumed to be zero in the present case. This resulted in much less satisfactory predictions for the structures in question (the predicted values were slightly better at the 99% level), and removal of these from the set led to a reduction in the sum of squares ratio between observed and predicted log (MBC) values from 0.27 to 0.13. The extent of the agreement between observed and predicted property values at this significance level is shown in Figure 36. Structures containing unique fragments are marked in parenthesis, and the 45° line, which would mark the correlation if predictions were completely accurate, is also indicated. The mean deviation between observed and predicted properties for the group which excludes the structures containing unique fragments is



0.45 log (MBC) units, compared with a range of 6.95 and a mean deviation for observed values of 1.43.

The results show the predictions to be in reasonably good agreement with observed property values, although as expected, they are not as good as the values estimated from the full regression analysis. The correlation between observed log (MBC) values and the values estimated from the full analysis at the 10% level is shown in Figure 37, and summaries of the property deviations in this case and at the 99% level are given in Table 12. The best estimated property values from the full regression are also listed with the best predictions in Table 13. These give a sum of squares ratio of  $< 0.01$  and a mean deviation between observed and estimated property values of 0.07 log (MBC) units.

The predictions for the 20 penicillins are summarised in Table 14. These are also reasonably good, and in this sample very similar results are obtained at the two different levels of significance, with the 99% level giving a better prediction in some cases. Regression coefficients obtained at higher confidence levels are in general expected to give more reliable estimates of the different substructural contributions to activity. In the present case however there are no detectable differences between the correlations at the 10% and 99% levels and the very close predictions obtained at these levels is not an unreasonable result in view of this. The different fragment types also perform very closely, and this is perhaps a more surprising result in view of the statistical differences indicated between the correlations obtained with these fragments. Octuplets, which give the best correlation result, give the least satisfactory predictions

at both significance levels, and augmented atoms and simple pairs give the best levels of prediction. Using the larger fragments more of the structures in the sample used for prediction contain unique fragment descriptions, but these structures, although less satisfactorily predicted in many cases, do not necessarily give rise to the worst predictions in the group. The slightly lower prediction levels given by bonded pairs and octuplets therefore cannot be satisfactorily explained on this basis. As mentioned above however the predictions are all very close and the particular variations arising in this small sample may not be of any practical significance. Possibly, investigations with larger samples, using larger design sets, in which all the substructures required for prediction are available at all times, would provide better indications of any important differences existing between fragments.

The best predictions, using augmented atoms at the 99% level are plotted against observed  $\log(B/F)$  values in Figure 38. In this case two of the structures contain unique fragment descriptions, and these are marked in parenthesis. They have been reasonably well predicted here, but give less satisfactory results in the bonded pair case. Other structures which are poorly predicted by each fragment contain substructures which have been excluded from the regression during the analysis of the design set. However, they cannot be explained on this basis alone, as such fragments are also present in some of the well predicted structures, and these are therefore more difficult to account for. The two positional isomers, structures 23 and 24, give identical predictions in each case. This is because they have identical observed serum binding measurements and under the par-

ticular conditions of the analysis they also have identical representations, i.e. none of the fragment types considered, including the larger definitions, are able to distinguish between them.

As in the previous sample the predicted property values, although reasonable, are not as good as the values estimated from the full structure set. The best agreement in this latter case, for the same sample of 20 structures, is shown in Figure 39, and the property deviations for each of the fragments used for prediction are summarised in Table 14. These give a lowest sum of squares ratio of 0.044 and a lowest mean deviation of 0.102 log (B/F) units, compared with lowest values of 0.126 and 0.187 log (B/F) units respectively in the prediction case (sample range 2.27, mean deviation for observed values 0.54). The estimated and predicted property values in these two cases are listed in Table 15.

#### 4.4.3 Interpretation of the Regression Solutions

In addition to giving approximate estimations of unknown property values, the regression coefficients obtained from the analyses should also give some indication of the influence of different substructures on the property in question. Except with atoms the substructural contributions obtained in each sample make sense chemically, and more detailed accounts of two of the analyses obtained at the 10% significance level are given below.

Table 16 gives the augmented pair results in the anaesthetics sample, and Table 17 the augmented atom contributions in the penicillins. In this second sample octuplets, which give the best overall explanation of the data, show a wider variety of chemical features in fragment definitions, and the regression coefficients in this case are much more difficult to interpret on a chemical basis. The medium

sized fragments are more straight-forward and augmented atoms in particular show up a number of important functional groups which enable closer comparisons to be made with other similar investigations reported in this area.

In the anaesthetics all except two of the regression coefficients are significant at the 1% level, and show the fragments containing carbon-carbon bonds tend to increase activity (negative coefficient values), whereas carbon-oxygen containing fragments in general decrease activity. Fragments containing carbon and tertiary nitrogen also tend to increase activity. However, those containing carbon and primary or secondary nitrogen bonds have coefficients which are not significantly different from zero at the 10% level, and these are not included in the regression. The chlorine containing fragment gives a negative coefficient, and its t statistic shows it to be a significant contributor, although it only occurs in one structure. The t-values listed show that the fragments with the highest coefficient values are not necessarily the most significant on statistical grounds, as this will depend on the way the fragments are distributed through the sample.

These results are consistent with the findings of Agin et al,<sup>209</sup> and of other authors<sup>219</sup> who report that local anaesthetic activity depends on the hydrophobic nature of the compound, with aromatic and other hydrophobic groups tending to increase activity and hydrogen-bonding groups to decrease it. To test whether the differences between coefficients were of statistical significance, pairs of coefficient values were compared using the following expression:

$$S_{(b_i, b_j)} = \sqrt{\frac{s_i^2 + s_j^2 - 2s^2 C_{ij}}{s}}$$

where  $S_i$  and  $S_j$  are the standard errors of the regression coefficients for fragments  $i$  and  $j$  respectively,  $s$  is the residual error of the regression and  $C_{ij}$  is the normalised cross-product term from the inverse cross-product matrix relating to fragments  $i$  and  $j$ . The significance level for the given fragment pair is then found by checking the value  $S(b_i, b_j)$  against values of Student's  $t$  distribution at the appropriate number of degrees of freedom, (i.e. the number of degrees of freedom for the given regression analysis).

During the comparison of coefficients particular attention was given to differences between similar substructures arising in chains and rings e.g. fragments 1C-C2 (chain) and 1C-C2 (ring), and 2C-01 (chain) and 2C-01 (ring) etc., and to differences between carbon-carbon chain fragments involving different degrees of substitution. None of the pairs examined, however, were found to differ significantly at the 5% level, or even at the 10% level. On statistical grounds, therefore, the different fragment contributions are equally significant and this means that the coefficient values cannot be taken as a measure of the relative importance of substructures. The individual contributions, however, can be regarded with some confidence as these are of high statistical significance and the different contributions are also largely in agreement with established trends.

Similarly in the penicillins the regression coefficients obtained are sensible from a chemical point of view, and agree with other recent investigations<sup>210,220</sup> reporting on the relationships between the hydrophobic nature of penicillin side chain structures and serum binding properties. Thus, results for augmented atoms in Table 17 show how the substructures containing hydrophilic groups, such as the

hydroxyl and the primary amine groups, tend to reduce serum binding (negative coefficient values), and fragments with hydrophobic properties such as the aromatic substructures, to increase binding. All except four of the coefficients are significant at the 1% level and two of these remaining four are significant at the 2% level. Once more, however, none of the coefficients differ significantly from each other at the 5% level, and the different contributions to serum binding must again be interpreted tentatively. The larger bond-centred fragments considered in this case also fail to show up any statistical differences between fragment contributions.

#### 4.5 Conclusions and Comparisons with other Regression Approaches on the same Data

The structure-property correlation method described here makes use of the technique of regression analysis and some techniques of substructural analysis to investigate a number of simple, empirical relationships between the structures and properties of organic compounds. The correlations obtained are very encouraging in view of the large approximations involved, and the two data sets used demonstrate the ability of the method to handle both related and dissimilar structural types.

Statistical tests were applied where possible to estimate the significance of the correlations and to test the differences arising between fragment performances. In cases where statistical significance is not indicated this does not mean that the results are not of some practical significance, although interpretations of the data must be more tentative in this case.

Highly significant correlations between structure and property data were obtained using a variety of substructural definitions. The strength of the relationship did not vary greatly with the type of descriptor, although in both samples, the larger fragments gave progressively better results, and, in some cases, correlations which were significantly better than those based on smaller definitions. No detectable differences were given between the correlations obtained at the 10% and 99% significance levels, and in the penicillins, the fragments giving the better correlations did not lead to better levels of prediction. This result may have something to do with the particular sample used for prediction, and the fact that many of the substructures required were either missing from the design set, or were excluded from it during the analysis. Other larger samples incorporating more of the substructures required for prediction may possibly show up wider differences, and would enable more reliable comparisons to be made.

Expressions involving quadratic terms did not lead to a significant improvement over a linear function, and where these expressions were considered they did not perform as well as a linear function based on larger fragments.

The smaller fragments led to the loss of some information on ring systems, and although the larger bond-centred fragments provided more detail of this nature, these were still unable to identify the different isomers possible, except for distinguishing ortho substituents from meta and para derivatives. Larger fragments could be generated automatically from connection table representations and this could provide more detailed information on ring systems. However,

such fragments involve a much wider variety of substructures, and the mathematical restrictions on the structure to feature ratio, would seriously limit their use. Ring information of this type would be better extracted from a linear notation, such as the Wiswesser Line Notation,<sup>221-223</sup> where explicit details of ring substituents and their location are provided.

The method compares well with the quantum-chemical and semi-empirical regression models described by Agin et al<sup>209</sup>, and Bird and Marshall,<sup>210</sup> and it has the advantage over these approaches of being more generally applicable, and of requiring fewer assumptions about the mode of action of compounds. The analysis obtained for the anaesthetics using Agin's expression is summarised in Table 10. Only two of the correlations based on structural descriptors differ significantly from this result, and a number of them give a slightly lower residual error. Figure 40 shows the agreement between observed and estimated log (MBC) values using Agin's expression, and Table 12 summarises the property deviations in this case. The mean deviation between observed and estimated property values is 0.18 log (MBC) units, and the sum of squares ratio 0.01, compared with lowest values of 0.07 log (MBC) units and < 0.01 in the case of structural descriptors. Similar results were obtained in the penicillins, and a summary of the analysis by Bird and Marshall is given in Table 11. As shown, a slightly lower residual error is given by a number of the larger fragment definitions, and the correlation based on octuplets differs significantly from the semi-empirical result at the 1% level. The agreement between observed and estimated log (B/F) values in the semi-empirical case, for example of 20 structures used for



prediction is shown in Figure 41, and the property deviations for these are summarised in Table 14. The smallest deviations using structural descriptors are again slightly better than those obtained by this method.

Property predictions by the semi-empirical approaches were not available for comparison, but the above very good comparisons, together with the reasonably good levels of prediction obtained by the 'hold-one-out' technique illustrate the possible value of the empirical approach for predicting biological activity.

The regression coefficients from the analyses could also be given a sensible explanation in terms of each substructure contribution to activity. Using each fragment the individual contributions to activity were reasonably highly significant in most cases, and although statistical differences were not indicated between substructures, the consistent results throughout and their close agreement with the physical interpretations given to regression equations in other similar investigations suggest the method could be of some help in rationalising the changes taking place in biological systems. However the method is very approximate, and several factors influence the significance of the regression coefficients. The fragments used are not independent of each other, and the overlapping substructures derived from the redundant connection table record add to the dependency problem. Smaller fragments involve less overlap, but very small substructures do not provide a satisfactory resolution of chemical types. This is better provided by larger fragments, but in addition to involving more overlap between substructures these eventually become more difficult to interpret chemically, as the functional groups con-

sidered important for activity become incorporated in larger sub-structures. The very approximate nature of substructural descriptors is another factor which limits the chemical interpretation of the regression equations. In view of these difficulties other more accurate quantum - mechanical constants, which are more independent and fundamental in nature, and can be calculated more specifically for different positions of the relevant molecules, may be better starting points for mode of action studies and providing information on the relative importance of different groups. However, even the best regression models eventually have to be tested using more direct experimental techniques, such as NMR and ESR methods, and from a practical point of view the above empirical method has the advantage over these of being very easily applicable and applicable on a wide scale in information systems which already hold structure - property files in suitable machine - readable form. It also has the important property of being able to handle structures which do not necessarily belong to the same chemical series, and has been shown here to handle these as effectively as closely related groups. This broadens the scope of the method considerably and in biological applications it enables the investigation of structures which do not belong to known active classes. Although approximate, the method could therefore be useful in preliminary drug design studies to point out compounds of potential biological interest before application of more accurate methods of analysis.

## CHAPTER 5

Discussion of the Classification and Regression Approaches  
as Methods for Structure - Property Correlation

The classification and regression methods described above compare very favourably with other similar structure-property correlation methods described in the recent literature, and the property predictions simulated in each case show that both hold promise as methods of prediction. The same set of structures and properties were used to compare the performances of the two approaches, and where direct comparisons were possible the classifications gave results which were comparable with the regressions.

The regression equations gave slightly better predictions on the whole, although not in every case. Those obtained for the anaesthetics and penicillin samples are summarised in Tables 18 and 19. In the anaesthetics the best regression equation, based on augmented pairs gave a sum of squares ratio between observed and predicted log (MBC) values of 0.27, compared with values of 0.46 and 0.48 in the classification and nearest neighbour cases using the same fragment definition. The best classification result in this sample is given by atom descriptions, which as shown give a better prediction level than the best regression result quoted above, both in the nearest neighbour and single-link cluster cases. A scatter diagram showing the slightly better agreement reached in the nearest neighbour case is given in Figure 42 (compare with the best regression result, Figure 36).

Predictions based on the best regression results in the penicillins, using augmented atoms, simple pairs, bonded pairs and octuplets are listed in Table 20 with the corresponding classification results for the same sample of 20 structures. Again, the results for each fragment are slightly better in the regression case. In this small sample very close performances are given by the two approaches. The best regression prediction is obtained using augmented atoms at the 99% significance level (sum of squares ratio 0.126), and the best classification result, based on augmented atoms and the simple distance coefficient is only slightly lower than this

(sum of squares ratio 0.181). Scatter diagrams showing the agreements reached in these two cases are given in Figures 38 and 43 respectively. The differences occurring in the anaesthetics are slightly wider, but it is unlikely that the variations in predictive performances arising in either sample are of statistical significance.

From these few investigations, therefore, it is difficult to make any definite statements on the relative performances of the two approaches. Some of the classification predictions are better than the regressions, but under similar conditions consistently better results are given by the more accurate regression method. These could well be of some significance, but because of the closeness of the results, it would be necessary to test this in other applications. With less accurate property measurements the interpretation of the regression solutions is more limited, and where only qualitative (nominal or ordinal) property measurements are available, the classification approach may be more appropriate. This point is discussed in Chapter 2. Regression analysis, however, may still be applied in these cases,<sup>172</sup> and although the individual substructural contributions to activity have less meaning here, it is possible that the regression equations will continue to give better overall agreements between structure and property data, and thus more accurate estimates of activity.

There are therefore two questions arising from this study which require further investigation. Firstly, it is necessary to establish whether the differences indicated in predictive performance are of some practical significance. Because of the very approximate nature of the structural descriptors used it may be that the regression approach has no particular advantage over the classification method, even when accurate property measurements are available. However, the regressions do give consistently better results, and if similar differences are observed in other appli-

cations this would increase the possible significance of these results. Secondly, if a difference is shown to exist between the two approaches, would similar relative performances be given with less accurate property measurements? Wider investigations with other types of property data would be needed to establish this.

The above additional information on relative performances would give a clearer indication of the most suitable roles which could be played by each approach in structure-property investigations. Depending on the performances found other factors could also influence the choice of methods. For example, the very useful statistical tests which can be applied in the regression case could be an important consideration. The majority of these tests are applicable even in the case of nominal property data, e.g. F and t tests are still valid in this case, and statistics such as the multiple correlation coefficient and residual error have their usual meaning. The regression coefficients can also be given a rough interpretation, although in this case it is not possible to compare them statistically. However, the sign and magnitude of these values still provide an approximate indication of which substructures are important contributors to the property in question and whether one contribution is more or less than another. This information and the various statistical criteria mentioned are not available in the classification case, but the classifications have a number of other useful properties which could be important in defining structure-property relationships. The method gives a rough pictorial representation of the data and shows approximately how the different structures under investigation are related to each other. This could be an extremely useful type of representation, for example, in showing up relationships between active and inactive derivatives. The classification method also has the advantage that much larger numbers of substructures

can be considered than in the regression case, although, as discussed in earlier chapters, much further research is required in this area to determine the exact changes brought about by altering the sample to feature ratio.

There are also computational differences to be considered. With the now widespread availability of standard statistical packages, the regression approach is much more easily applicable. However, the classification method has a potentially very useful application in large computer-based files for the storage and retrieval of chemical structure information, and if a practical application is found in this area, it would mean that the data would already be in a form suitable for structure-property calculations of this type

Each approach, therefore, has a number of properties in its favour, and as they both present the data differently it may be that the choice of methods, where a choice exists, will depend on the type of application in question. From a prediction point of view, the odds are weighed slightly in favour of the regression method, and unless considerable differences in the relative performances of the two approaches are indicated in other applications, it is expected that this will be the preferred approach in applications where structure-property correlation and prediction are the main objectives. However, this would not rule out the very useful contributions which could be made by the classification method in preliminary studies, and these could be used to show up important structural relationships in the data and to give rough property assignments before the application of more accurate methods.

DESCRIPTION OF COMPUTER PROGRAMS



Computation was carried out on the Sheffield University ICL 1907 computer, which has a 24-bit word length and a cycle time of approximately 2  $\mu$ s.

#### i Data Files

Structures were coded as redundant connection tables on punched cards initially, and these were used to store the connection tables on magnetic tape or disc by user program. The redundant records were based on a compacted, multi-level description, details of which have been reported elsewhere.<sup>224</sup> In them each bond is specified twice, once at each atom it links, and in addition to giving bond orders the record also indicates whether bonds are present in rings or chains. An example of the redundant record is given in Figure 44.

#### ii Software

Computer programs were written in PLAN (the ICL assembly language), FORTRAN and ALGOL. Only limited facilities were available at the outset of these investigations for the manipulation of source programs on magnetic tape or disc, particularly in the case of PLAN programs, and because of the numerous modifications required during the course of the study, user programs were therefore retained on card files.

#### ii(a) The Classification Programs

The main classification programs were written in PLAN, and these incorporated PLAN and FORTRAN subroutines for the calculation of similarity and dissimilarity coefficients. Initially, connection tables were analysed and the different structural features present listed. These were then used to derive the appropriate numerical representations, upon which the calculation of SCs and DCs were based. The setting up of numerical representations and comparison of structures were carried out in different segments of the same program and the resulting SC or DC values were stored on magnetic tape ready

for the clustering phase. The central processing times and core storage requirements of this first phase varied slightly with the sample size and the type of association measure used (core storage between 2,000 and 4,000 words + working storage: CPU times ranging from a few seconds in the aminoacids sample to up to 50-60 seconds in the two larger samples).

Before clustering, a ranked listing of similarity and dissimilarity coefficients was required. The particular single-link clustering algorithm used generates clusters level by level, starting at the highest level, and this meant it was necessary to arrange coefficients either in decreasing order of similarity or increasing order of dissimilarity. This step was carried out on magnetic tape, using standard ICL sorting routines before the coefficients were input to the clustering program. Prior to clustering it was also necessary to identify the structure pair associated with each coefficient value, to note all the pairs arising at each different level and finally to determine the maximum number of pairs arising with a given similarity coefficient, so that the required arrays could be set for clustering. These various tasks were performed by a PLAN program, which incorporated the FORTRAN clustering algorithm in the form of a subroutine. The calling program was also required to initialise count fields, and to zeroise the arrays in which the cluster information is produced. The clustering routines were called for each new value of similarity or dissimilarity, after all the structure pairs arising at the level in question had been placed in the appropriate pair vectors. A listing of the clusters formed at each level was produced, except where these are identical with those arising at the previous level. Single-element clusters were ignored. Variations in central processing times with the different sample size were not as wide in this second phase, and core storage requirements were roughly the same (CPU times ~ 20 seconds: core storage between 3,000 and 5,000 words

+ working storage).

The basic clustering routines were modified slightly to identify the highest associations between structures, and the associations arising for each structure in the cluster it first joins. These were the pair values required for prediction, and they were listed by the calling program before being output to magnetic tape in readiness for the prediction phase.

Prediction programs were written in ALGOL. The pair values together with observed property values, read in separately from a card file, were used to estimate a property value for each structure, firstly in the nearest neighbour case, then in the classification. Any number of predictions could be carried out in the same computer run, provided these were specified on input. A listing was produced of predicted property values, deviations between observed and predicted values, mean property deviations, and other statistical quantities important for estimating the agreement between observed and predicted values, such as variance ratios and standard deviations.

The classification programs were linked by means of job control statements so that they could be run in series, and the standard software routines used were also linked in with these so that all operations could be carried out in the same computer run. A flowchart of the basic operations is given in Figure 45. Only the important input/output operations have been indicated. Nearest neighbours and first cluster associations output at the end of the second stage were not in a form suitable for input to the ALGOL prediction routines. The record layouts produced by the PLAN 'write' statements during the second stage were preceded by a word count field and in this format they could not be read directly by standard ALGOL 'read' routines. To interface the prediction and clustering phases, therefore, a small PLAN subprogram was used, which first of all extracted the appropriate pair values from the

magnetic tape record, and secondly converted these from fixed point to floating point form, suitable for ALGOL processing. The PLAN 'decoding' statements were held as a special subroutine and were called by the prediction programs each time a new pair of structure values was required.

ii(b) The Regression Programs

The regression analyses were carried out using the ICL statistical analysis package<sup>219</sup>, and supporting programs to obtain the data in a form suitable for input to the statistical package were written in PLAN and FORTRAN. Prediction routines were written partly in PLAN and partly in ALGOL.

The frequency vectors required for regression, giving the frequency of occurrence of the different fragment types arising in each structure, were very similar to the representations required in the classification case for the calculation of distance functions and the quantitative probabilistic measures, and almost identical routines were employed here for the analysis of connection tables. In this case the vectors were used to set up a so-called observation matrix, consisting of the fragment frequency values, designated as the set of independent variables, a structure identification field, and a property value, which was read in separately and identified as the dependent variable on input to the regression package (see Figure 33). Where necessary the PLAN programs developed to generate the observation matrix incorporated FORTRAN subroutines for the derivation of the appropriate quadratic terms. Detailed descriptions of the type of input formats required by the regression package are given in the ICL manual.

The regression solutions and regression coefficients were listed on the line-printer and other useful information was produced, such as details of the matrices used during the course of the analysis, and details of estimated property values and their deviations from observed values. This additional

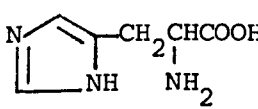
information was requested on input.

It was not possible to interface the regression package with user programs, and this meant that the matrix generation, regression analysis and prediction phases could not be run in series, in the same way as the classification programs. Another limitation of the package in the system in question was that input was required in card form. The regression coefficients needed for prediction were also output in this form which meant prediction involved three separate stages altogether. Firstly, in the data generation phase, the appropriate frequency vectors for the structures undergoing prediction were excluded from the observation matrix, one at a time. The regression coefficients produced by these reduced matrices, together with observed property values, were then returned to the PLAN program, used to generate the original data matrix, for calculation of the appropriate property value. Predictions were batched to save time. The resulting sets of observed and predicted property values were then input to the ALGOL routines used for prediction in the classification case, to determine the extent of the agreement between observed and estimated values. A summary of the different stages involved is given in Figure 46.

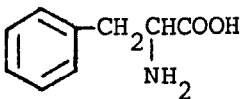
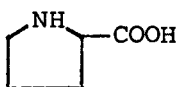
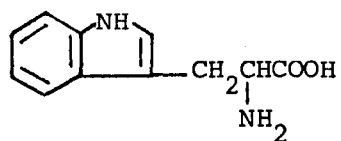
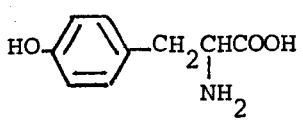
APPENDICES

I Data Sets and Properties

Sample 1 : 20 Naturally Occurring Amino Acids <sup>10</sup>

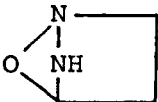
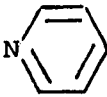

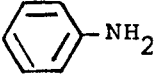
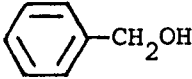
	<u>Structure</u>	<u>pI Value</u>
1	$\begin{array}{c} \text{CH}_3\text{CHCOOH} \\   \\ \text{NH}_2 \end{array} \quad (\text{alanine})$	6.00
2	$\begin{array}{c} \text{HN}=\text{CNHCH}_2\text{CH}_2\text{CH}_2\text{CHCOOH} \\   \qquad \qquad \qquad   \\ \text{NH}_2 \qquad \qquad \qquad \text{NH}_2 \end{array} \quad (\text{arginine})$	10.76
3	$\begin{array}{c} \text{HOOCCH}_2\text{CHCOOH} \\   \\ \text{NH}_2 \end{array} \quad (\text{aspartic acid})$	2.77
4	$\begin{array}{c} \text{H}_2\text{NCCH}_2\text{CHCOOH} \\    \qquad   \\ \text{O} \qquad \text{NH}_2 \end{array} \quad (\text{asparagine})$	5.41
5	$\begin{array}{c} \text{HSCH}_2\text{CHCOOH} \\   \\ \text{NH}_2 \end{array} \quad (\text{cysteine})$	5.07
6	$\begin{array}{c} \text{HOOCCH}_2\text{CH}_2\text{CHCOOH} \\   \\ \text{NH}_2 \end{array} \quad (\text{glutamic acid})$	3.22
7	$\begin{array}{c} \text{H}_2\text{NCCH}_2\text{CH}_2\text{CHCOOH} \\    \qquad   \\ \text{O} \qquad \text{NH}_2 \end{array} \quad (\text{glutamine})$	5.65
8	$\text{H}_2\text{NCH}_2\text{COOH} \quad (\text{glycine})$	5.97
9	 $\begin{array}{c} \text{N} \quad \text{CH}_2\text{CHCOOH} \\ \diagdown \quad   \\ \text{CH} \quad   \\ \diagup \quad \text{NH} \quad   \\ \text{NH}_2 \end{array} \quad (\text{histidine})$	7.59
10	$\begin{array}{c} \text{CH}_3\text{CH}_2\text{CH}-\text{CHCOOH} \\   \qquad   \\ \text{CH}_3 \quad \text{NH}_2 \end{array} \quad (\text{isoleucine})$	6.02
11	$\begin{array}{c} \text{CH}_3\text{CHCH}_2\text{CHCOOH} \\   \qquad   \\ \text{CH}_3 \quad \text{NH}_2 \end{array} \quad (\text{leucine})$	5.98

(20 Naturally Occurring Amino Acids continued)

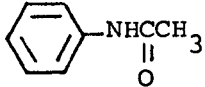
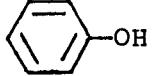
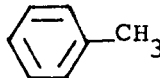
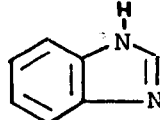
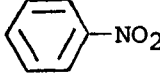
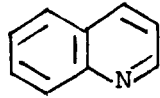
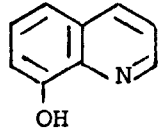
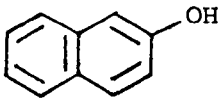
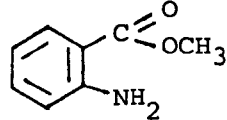
	<u>Structure</u>	<u>pI Value</u>
12	$\text{H}_2\text{NCH}_2\text{CH}_2\text{CH}_2\text{CH}_2\text{CH}(\text{NH}_2)\text{COOH}$ (lysine)	9.74
13	$\text{CH}_3\text{SCH}_2\text{CH}_2\text{CH}(\text{NH}_2)\text{COOH}$ (methionine)	5.74
14	 (phenylalanine)	5.48
15	 (proline)	6.30
16	$\text{HOCH}_2\text{CH}(\text{NH}_2)\text{COOH}$ (serine)	5.68
17	$\text{HOCH}(\text{CH}_3)\text{CH}(\text{NH}_2)\text{COOH}$ (threonine)	6.16
18	 (tryptophan)	5.89
19	 (tyrosine)	5.66
20	$\text{CH}_3\text{CH}(\text{CH}_3)\text{CH}(\text{NH}_2)\text{COOH}$ (valine)	5.96



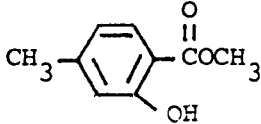
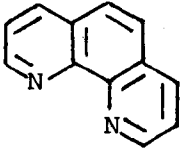
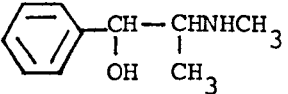
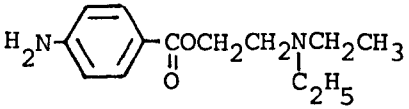
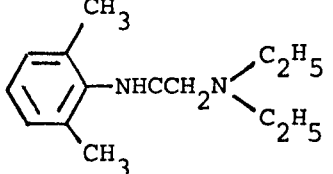
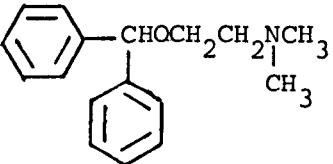
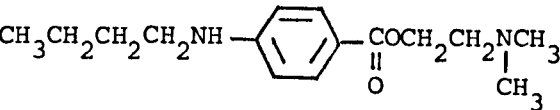
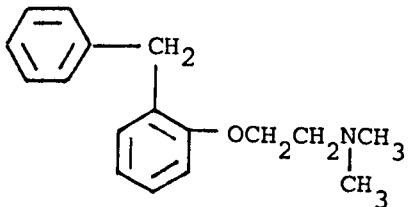
Sample 2: 39 Local Anaesthetics 209

	<u>Structure</u>	<u>Log (MBC) Value</u>
1	CH <sub>3</sub> OH (methanol)	3.09
2	CH <sub>3</sub> CH <sub>2</sub> OH (ethanol)	2.75
3	$\begin{array}{c} \text{CH}_3\text{CCH}_3 \\    \\ \text{O} \end{array}$ (acetone)	2.60
4	$\begin{array}{c} \text{CH}_3\text{CHCH}_3 \\   \\ \text{OH} \end{array}$ (isopropanol)	2.55
5	CH <sub>3</sub> CH <sub>2</sub> CH <sub>2</sub> OH (propanol)	2.40
6	$\begin{array}{c} \text{N}=\text{CCH}_2\text{NHCOCH}_2\text{CH}_3 \\    \\ \text{O} \end{array}$ (urethane)	2.00
7	CH <sub>3</sub> CH <sub>2</sub> OCH <sub>2</sub> CH <sub>3</sub> (ethyl ether)	1.93
8	CH <sub>3</sub> CH <sub>2</sub> CH <sub>2</sub> CH <sub>2</sub> OH (butanol)	1.78
9	 (antipyrine)	1.78
10	 (pyridine)	1.77
11	CHCl <sub>3</sub> (chloroform)	1.50
12	 (hydroquinone)	1.40
13	 (aniline)	1.30
14	 (benzylalcohol)	1.30

(39 Local Anaesthetics continued)

	<u>Structure</u>		<u>Log(MBC) Value</u>
15	 (acetanilide)		1.17
16	CH <sub>3</sub> CH <sub>2</sub> CH <sub>2</sub> CH <sub>2</sub> CH <sub>2</sub> OH (pentanol)		1.20
17	 (phenol)		1.00
18	 (toluene)		1.00
19	 (benzimidazole)		0.81
20	CH <sub>3</sub> CH <sub>2</sub> CH <sub>2</sub> CH <sub>2</sub> CH <sub>2</sub> CH <sub>2</sub> OH (hexanol)		0.56
21	 (nitrobenzene)		0.47
22	 (quinoline)		0.30
23	 (8-hydroxyquinoline)		0.30
24	CH <sub>3</sub> CH <sub>2</sub> CH <sub>2</sub> CH <sub>2</sub> CH <sub>2</sub> CH <sub>2</sub> CH <sub>2</sub> OH (heptanol)		0.20
25	 (2-naphthol)		0.00
26	 (methylantranilate)		0.00
27	CH <sub>3</sub> CH <sub>2</sub> CH <sub>2</sub> CH <sub>2</sub> CH <sub>2</sub> CH <sub>2</sub> CH <sub>2</sub> CH <sub>2</sub> OH (octanol)		-0.16

(39 Local Anaesthetics continued)

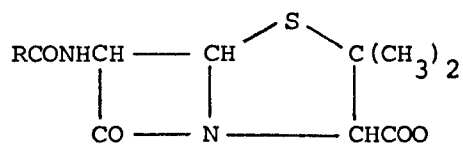
	<u>Structure</u>	<u>Log(MBC) Value</u>
28	 <p>(thymol)</p>	-0.52
29	 <p>(O-phenanthroline)</p>	-0.80
30	 <p>(ephedrine)</p>	-0.80
31	 <p>(procaine)</p>	-1.67
32	 <p>(xylocaine)</p>	-1.96
33	 <p>(diphenhydramine)</p>	-2.80
34	 <p>(tetracaine)</p>	-2.90
35	 <p>(phenyltoloxamine)</p>	-3.20

(39 Local Anaesthetics continued)

	<u>Structure</u>		<u>Log(MBC) Value</u>
36		(quinine)	-3.60
37		(eserine)	-3.66
38		(caramiphen)	-4.00
39		(dibucaine)	-4.20

Sample 3: 79 Penicillins 210

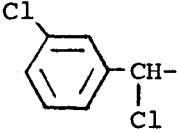
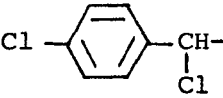
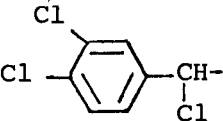
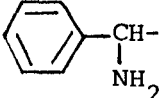
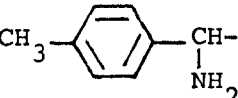
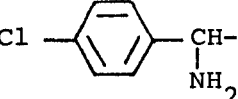
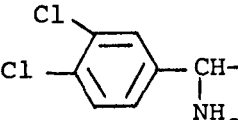
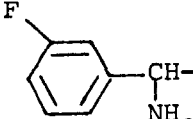
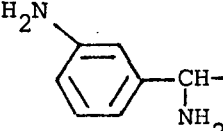
Parent Compound



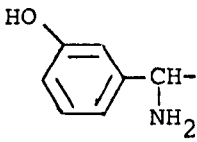
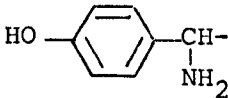
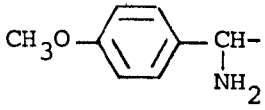
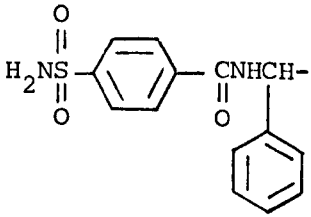
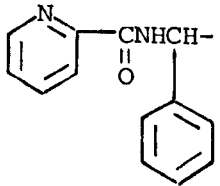
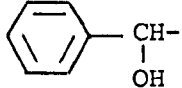
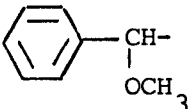
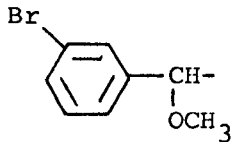
<u>Penicillin</u>	<u>R</u>	<u>Log(B/F) Value</u>
1*	H-	-0.659
2	CH <sub>3</sub>	-0.753
3	CH <sub>3</sub> CH <sub>2</sub> CH <sub>2</sub> CH <sub>2</sub> CH <sub>2</sub> CH <sub>2</sub> CH <sub>2</sub> -	1.085
4*	$  \begin{array}{c}  \text{C}_3\text{H}_7 \\    \\  \text{C}_3\text{H}_7\text{C}- \\    \\  \text{C}_3\text{H}_7  \end{array}  $	1.144
5	CH <sub>3</sub> OCH <sub>2</sub> -	-1.110
6	CH <sub>3</sub> CH <sub>2</sub> OCH <sub>2</sub> -	-0.410
7	CH <sub>3</sub> CH <sub>2</sub> CH <sub>2</sub> CH <sub>2</sub> OCH <sub>2</sub> -	0.154
8*	$  \begin{array}{c}  \text{CH}_3\text{CH}_2\text{CHOCH}_2- \\    \\  \text{CH}_3  \end{array}  $	-0.052
9*	$  \begin{array}{c}  \text{CH}_3\text{CH}_2\text{CH}- \\    \\  \text{OCH}_3  \end{array}  $	-0.602
10	$  \begin{array}{c}  \text{CH}_3\text{CH}_2\text{CH}_2\text{CH}_2\text{CH}- \\    \\  \text{OCH}_2\text{CH}_3  \end{array}  $	0.454
11	CH <sub>3</sub> CH <sub>2</sub> OCH <sub>2</sub> CH <sub>2</sub> -	-0.477
12	$  \begin{array}{c}  \text{CH}_3\text{CH}_2\text{CH}_2\text{CH}- \\    \\  \text{NH}_2  \end{array}  $	-0.308



(79 Penicillins continued)

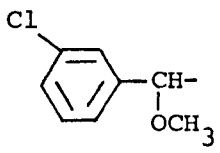
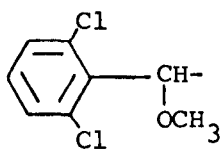
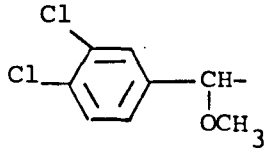
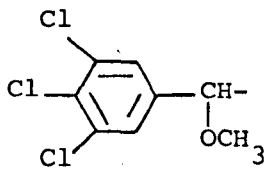
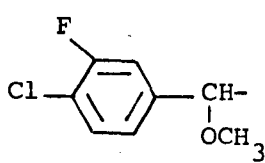
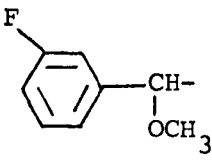
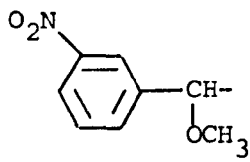
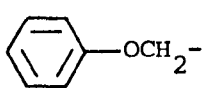
<u>Penicillin</u>	<u>R</u>	<u>Log(B/F) Value</u>
23*		1.195
24*		1.195
25		1.510
26		-0.659
27*		0.176
28		0.087
29*		0.664
30		-0.454
31		-0.865

(79 Penicillins continued)

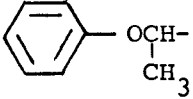
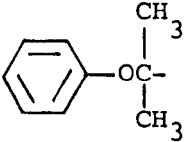
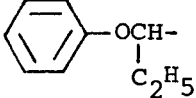
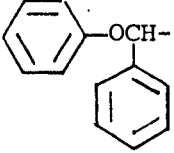
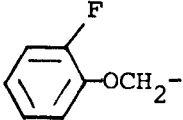
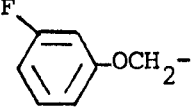
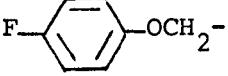
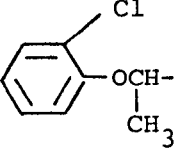
<u>Penicillin</u>	<u>R</u>	<u>Log(B/F) Value</u>
32*		-0.695
33		-0.575
34		-0.213
35		-0.140
36		0.231
37		0.056
38		0.213
39		0.865



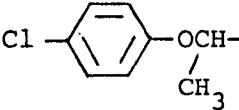
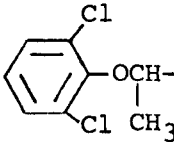
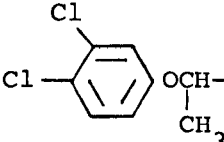
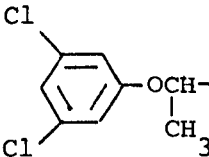
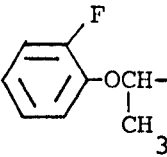
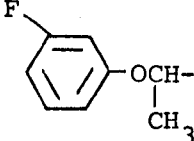
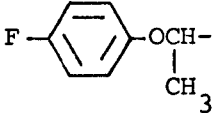
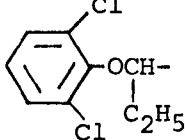
(79 Penicillins continued)

<u>Penicillin</u>	<u>R</u>	<u>Log(B/F) Value</u>
40		0.689
41		0.720
42		1.061
43		1.440
44		0.689
45		0.269
46*		0.176
47		0.589

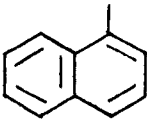
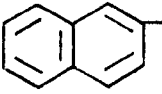
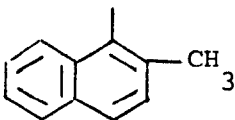
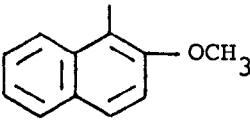
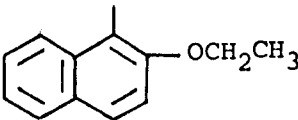
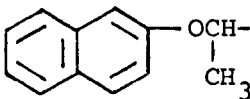
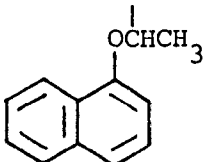
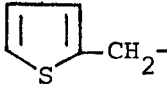
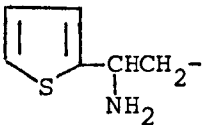
(79 Penicillins continued)

<u>Penicillin</u>	<u>R</u>	<u>Log(B/F) Value</u>
48*		0.644
49		1.091
50		0.792
51		1.541
52		1.032
53		0.704
54		0.644
55*		1.380

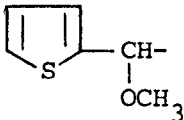
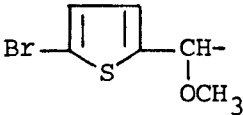
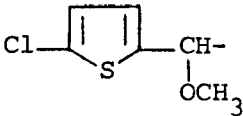
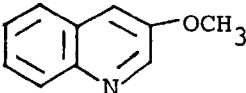
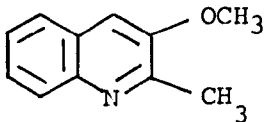
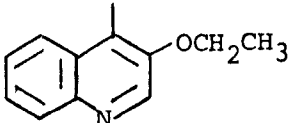
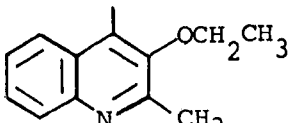
(79 Penicillins continued)

<u>Penicillin</u>	<u>R</u>	<u>Log(B/F) Value</u>
56*		1.261
57		1.380
58		1.574
59		1.510
60		0.788
61		0.720
62		0.602
63		1.297

(79 Penicillins continued)

<u>Penicillin</u>	<u>R</u>	<u>Log(B/F) Value</u>
64		0.788
65		1.337
66		0.661
67*		0.602
68*		0.921
69		1.252
70*		1.574
71		0.140
72		-0.327

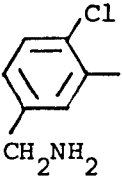
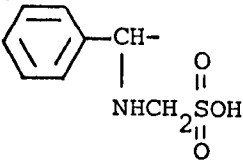
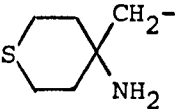
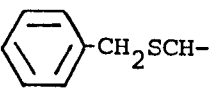
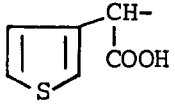
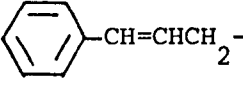
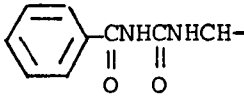
(79 Penicillins continued)

<u>Penicillin</u>	<u>R</u>	<u>Log(B/F) Value</u>
73		0.158
74		0.940
75		0.707
76*		0.122
77		0.207
78*		0.362
79*		0.466

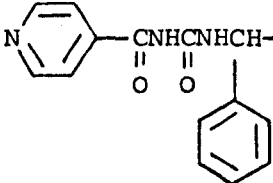
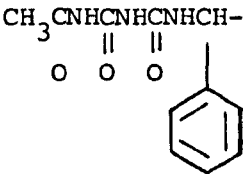
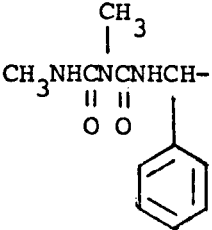
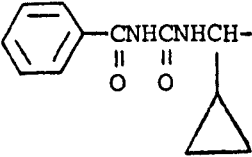
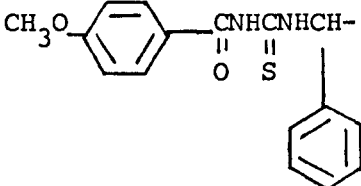
\* Penicillins 'predicted' by the regression method using the 'hold-one-out' technique



(18 Penicillins continued)

<u>Penicillin</u>	<u>R</u>	<u>B Value</u>
7		67
8		11
9		5
10		49
11		45
12		88
13		94

(18 Penicillins continued)

<u>Penicillin</u>	<u>R</u>	<u>B Value</u>
14		50
15		26
16		50
17		86
18		95



## II Altering the Sample to Feature Ratio in Classification Applications

In designing a classification system one of the most serious problems arising as discussed earlier is deciding on the number of features to be used when a finite number of objects is available. Some attention has been given to this question in the present investigation, although the statistical problems arising have only been touched upon very briefly. The effects of altering the sample to feature ratio is an area which has been considered recently by a number of investigators, but so far investigations have been restricted to very simple two-way classification systems, where each class has equal 'a priori' probabilities, and it has been easy to measure the experimental classification result against expected theoretical performances. In such simple systems it has been shown that the error rate on the design data is a monotonically increasing function of the ratio of sample size for feature size, and that quite wide discrepancies between observed and expected error rates arise when the sample to feature ratio falls below 3. In real classification situations, which are much more complicated than the above system and where less is known about the probability structure, the problem is much more difficult to evaluate in terms of observed and expected classification performances. Unsupervised systems are even more difficult to assess, because there is no prior knowledge of class structure in this case, and a formal approach is virtually impossible. Until the properties of these systems are understood more fully, therefore, it is only possible to tackle the problem empirically, by examining method performances in relation to sample to feature ratios.

In the present study a few empirical investigations were carried out on a number of random samples taken from the group of 79 penicillins. Progressively larger samples were considered, and it was found that the

resulting increase in the structure to feature ratio did not lead to a noticeable improvement in predictive performance. The samples used, however, were unsatisfactory for a number of reasons. Firstly, they did not lead to very wide differences in the structure to feature ratio. Secondly, because a different sample was considered in each case, direct comparisons of the different predictive performances were not strictly valid. A more satisfactory approach would be to vary the number of sub-structures in a fixed sample, rather than varying the sample size, but additional investigations along these lines were not possible in the time available.

Further work is therefore needed in this area, and it is hoped that the increased attention given to statistical problems in supervised learning systems will encourage a similar interest in the unsupervised case, as more serious investigations are carried out with this type of approach.

### III Semi-empirical Structure-Property Correlations using Structural Parameters

The very good agreement between structure and property data using the above empirical regression model led to some additional investigations in this area to see whether the methods of handling chemical structures automatically could be equally effective in the case of semi-empirical problems. Time and data limitations prevented a very thorough investigation of the area, but some useful studies were possible which demonstrated the potential value of the approach.

In the method developed, the property parameters derived from the analysis of one set of structures were used to predict the properties of another set, and estimated properties were subsequently correlated with some observed property. The approach compares closely with the semi-empirical correlation method developed by Hansch and co-workers, except structure-property relationships in this case are based on the structural features of the entire molecule. The group of 79 penicillin structures used to test the empirical regression model were considered again here, so that direct comparisons between the two approaches could be made. Using these, the primary objective was to establish whether useful correlations could be obtained between observed serum binding measurements and estimated partition coefficients, where the latter are calculated from the fragment  $\Pi$  contributions derived from an independent set of structures, analysed by the empirical method. This second structure set, for which partition coefficients were already available, was referred to as the training or learning set, and the required fragment  $\Pi$  values were obtained from its analysis using the following expression:

$$1 \quad \log P_i = \sum_{j=1}^n \Pi_j x_{ij} + \text{const}$$

where  $\log P_i$  is the observed partition coefficient for learning set structure  $i$ , and the quantities  $n$ ,  $\Pi_j$ ,  $X_{ij}$  and  $\text{const}$  are as defined in Chapter 4 ( $\Pi_j$  was defined previously as  $b_j$  - the regression coefficient derived from the analysis for fragment  $j$ ). Using the appropriate fragment  $\Pi$  values, partition coefficients were then estimated for the penicillins as follows:

$$2 \quad \log P_k = \sum_{j=1}^n \Pi_j X_{ij} + \text{const}$$

where there are  $m$  fragment types in the sample,  $X_{kj}$  is the frequency of occurrence of fragment  $j$  in penicillin  $k$ ,  $\Pi_j$  is the regression coefficient derived from the learning set for fragment  $j$  and  $\text{const}$  is the regression constant derived from the same analysis. The empirical regression model developed by Nys and Rekker<sup>214</sup> is similar to this, except for the differences in structural description outlined in Chapter 4.

Observed serum binding values were finally correlated with estimated partition coefficients, using the observed property as the dependent variable as follows:

$$3 \quad \log (B/F)_{\text{observed}} = a \log P_{\text{calculated}} + c$$

where  $a$  is the regression coefficient from the analysis, and  $c$  the regression constant.

One of the difficulties of this approach is obtaining suitable learning set structures with the required property information. The required structures in the present case, i.e. structures with suitable substructures and known partition coefficients, were obtained from a number of different

literature sources, and one important feature of the resulting learning set was that it contained quite a wide variety of structural types, ranging from simple chain structures to some simple benzenoid and fused heterocyclic derivatives, none of which were related closely to the penicillins. Because of this, difficulties were encountered in obtaining some of the larger penicillin substructures, and it was necessary to restrict investigations to smaller fragments which could account for most of the substructures present in the test set. Of the fragment types considered simple pairs were able to account for all the penicillin substructures, and some of the results obtained with these are reported below.

The learning set, consisting of 130 structures in all, contained 26 simple pair fragments, 17 of which arose in the group of 79 penicillins. Table 20 summarises the results of the regression analyses carried out on the learning set at the 99% and 10% significance levels using these fragments as the independent variables. A significant difference between the two levels was not indicated, and both correlations were found to be significant at the 0.1% level. At the 99% level of analysis all of the substructures required for the prediction of the penicillins were included in the regression set. At the 10% level some of the less significant ones were excluded, and just as in the empirical model, the missing fragment

values in this case were assumed to be zero. The parameters obtained at this level resulted in slightly less satisfactory partition coefficient estimations in the penicillin sample, and the different agreements reached with observed serum binding values in the two cases are summarised in Table 21. The slightly better property estimations, however, did not lead to a significant improvement in the correlation, and both results were significant at the 0.1% level.

In view of the good correlation obtained in the penicillin sample, the investigations were taken a stage further, and an attempt was made to use both the original learning set and the penicillins, as learning sets to predict serum binding values in another smaller group of penicillins, where neither property was assumed available. Details of the sample of 18 penicillin structures used in this investigation are given in Appendix 1. Partition coefficients were first estimated for the group, as described above, and these quantities were then substituted in the right hand side of expression 3, to predict a serum binding value for each structure, using the slope and intercept values derived from the analysis of the larger penicillin sample.

Due to greater variations in side chain structures in this sample, a wider range of substructures was present, and not all of these were present in the learning set of 130 structures used to predict the larger sample. Suitable extensions to this learning set were not possible in the time available, and the three additional simple pair fragments arising in the sample were assumed to have zero  $\pi$  contributions. The structures containing these fragments were therefore expected to be less well predicted. Estimated partition coefficients were then used in expression 3 to obtain  $\log (B/F)$  values, using the regression constant and coefficient values obtained from the analysis of the larger sample which gave the lowest residual error.

$$4. \quad \log (B/F)_{\text{calc}} = 0.6619 \log P_{\text{calc}} + 1.1726$$

The best  $\log (B/F)$  predictions were obtained using the partition coefficients derived from fragment  $\pi$  contributions obtained at the 99% level, and the corresponding B values for these are listed with observed values in Table 22. Compounds containing the fragments not present in the learn-

ing set of 130 structures are asterisked. As expected, most of these were poorly predicted. Some other structures were also poorly predicted, for example structures 10, 16 and 17. The quite wide discrepancy between observed and predicted serum binding values in structure 10 is an interesting result, as the higher predicted value is much nearer the expected value for this structure on a hydrophobicity basis. The other wider differences arose in the case of the multiple amide structures 13 to 17, but it has been noted that these structures are also difficult to handle by the Hansch method.<sup>224</sup> Remaining structures were reasonably predicted, and a few were very well predicted such as the thiacyclohexane (9) and the phoxymethyl (6) derivatives. The carboxy compounds 1, 2 and 11 were also quite well predicted.

The results obtained in these few investigations were extremely encouraging in view of the limitations of the learning sets and the numerous approximations involved. Correlation coefficients were statistically significant, although smaller than usually obtained in empirical investigations under similar conditions. Log (B/F) predictions were also reasonably good in the smaller test sample, considering the two different learning sets involved in this case, one of which did not contain all the fragment types required. Many other fragments needed for prediction were present in only a few learning set structures, and an additional problem in the learning set used to predict partition coefficients was the number of dissimilar structural types involved. This meant that the fragment  $\Pi$  contributions derived from this set may not have been entirely appropriate for prediction of the two penicillin groups, because of environmental differences between samples. Investigations could not be taken any further in the time available, but it is expected that larger, more representative learning sets, allowing for the investigation of a wider variety of substructures, will

lead to improvements in the correlations, and, hopefully, predictions which are comparable with those obtained in the empirical case.



## BIBLIOGRAPHY

- 1 Salton, G. "Automatic Information Organisation and Retrieval" McGraw-Hill, New York (1968). "The SMART Retrieval System - Experiments in Automatic Document Processing." Prentice-Hall, Englewood Cliffs (1971).
- 2 Jardine, N. and van Rijsbergen, C.J. "The Use of Hierarchic Clustering in Information Retrieval." Information Storage and Retrieval, 7, 217-240 (1971).
- 3 Spark Jones, K. "Automatic Keyword Classification for Information Retrieval." Butterworth, London (1971).
- 4 Hansch, C. "Quantitative Structure-Activity Relationships in Drug Design," in: Drug Design I, 271-342 ed. E.J. Ariens, Medicinal Chemistry Series, 11, Academic Press, New York (1971).
- 5 Redl, G., Cramer, R.D.III and Berkoff, C.E. "Quantitative Drug Design" Chem. Soc. Rev., 28, 273-292 (1974).
- 6 Cammarata, A. "Quantitative Structure - Activity Relationships." Annu. Rep. Med. Chem., Section IV, 245-253 (1971).
- 7 Free, S.M. Jr. and Wilson, J.W. "A Mathematical Contribution to Structure-Activity Studies." J. Med Chem., 7, 395-399 (1964).
- 8 Verloop, A. "The Use of Linear Free Energy Parameters and Other Experimental Constants in Structure-Activity Studies," in: Drug Design III, 133-187 ed. E.J. Ariens, Academic Press, New York, London (1972).
- 9 Goodford, P.J. "Prediction of Pharmacological Activity by the Method of Physicochemical-Activity Relationships," in: Advances in Pharmacology and Chemotherapy II, 51-97 (1973).
- 10 Sneath, P.H.A. "Relations between Chemical Structure and Biological Activity in Peptides." J. Theoret. Biol., 12, 157-195 (1966).
- 11 Koskinen, J.R. and Kowalski, B.R. "Structure-Activity Correlations for Organic Molecules by Pattern Recognition," NTIS Report AD-785 913 (1974).
- 12 Kowalski, B.R., and Bender, C.F. "The Application of Pattern Recognition to Screening Prospective Anticancer Drugs. Adenocarcinoma 755 Biological Activity Test." J. Amer. Chem. Soc., 96 (3), 916-918 (1974).

- 13 Stuper, A.J. and Jurs, P.C. "Classification of Psychotropic Drugs as Sedatives or Tranquilizers using Pattern Recognition Techniques." J. Amer. Chem. Soc., 97(1), 182-187 (1975).
- 14 Palmer, G. "Wiswesser Line Formula Notation." Chem. Brit., 6 422-426 (1970).
- 15 Lynch, M.F., Harrison, J.M., Town, W.G. and Ash, J.E. "Computer Handling of Chemical Structure Information." Macdonald, London (1971).
- 16 Granito, C.E. and Garfield, E. "Substructure Search and Correlation in the Management of Chemical Information." Naturwissenschaften, 60, 189-197 (1973).
- 17 Adamson, G.W., Corwell, J., Lynch, M.F., Mclure, A.H.W., Town, W.G. and Yapp, A.M. "Strategic Considerations in the Design of a Screening System for Substructure Searches of Chemical Structure Files." J. Chem. Documentation, 13 (3) 153-157 (1973).
- 18 Van Rijsbergen, C.J. "An Algorithm for Information Structuring and Retrieval." The Computer Journal, 14 (4), 407-412 (1971).
- 19 Lachenbruch, P.A. "Estimation of Error Rates in Discriminant Analysis." Ph.D. dissertation Univ. Southern California, Los Angeles (1965).
- 20 Kowalski, B.R. and Bender, C.F. "Pattern Recognition. A Powerful Approach to Interpreting Chemical Data." J. Amer. Chem. Soc., 94 (16) 5632-5639 (1972).
- 21 Sneath, P.H.A. "Some Statistical Problems in Numerical Taxonomy." The Statistician, 17 (1), 1-12 (1967).
- 22 Bond, V.B., Bowman, C.M., Lee, N.L., Peterson, D.R. and Reslock, M.H. "Interactive Searching of a Structure and Biological Activity File." J. Chem. Documentation, 11, 168-170 (1971).
- 23 Hyde, E., Lambourne, D.R. and McArdle, L.A. Abstracts of Papers, 163rd. National Meeting of the ACS Boston, April (1972).
- 24 Jacobus, D.P., Davidson, D.E., Feldman, A.P. and Schafer, J.A. "Experience with the Mechanized Chemical and Biological Information Retrieval System." J. Chem. Documentation, 10, 135-140 (1970).
- 25 Heincke, F. "Naturgeschichte des Herings. I. Die Lokalformen und die Wanderungen des Herings in den europäischen Meeren." Abb. Deutsch. Seefischerei-Vereins, 2, i-cxxxxvi, 1-223 (1898).

- 26 Czekanowski, J. "Zur Differentialdiagnose der Neandertalgruppe." Korrespondenzblatt Duetsch. Ges. Anthropol. Ethnol. Urgesch., 40, 44-47 (1909).
- 27 Pearson, K. "On the Coefficient of Racial likeness." Biometrika, 18, 105-117 (1926).
- 28 Rao, C.R. "The Utilization of Multiple Measurements in Problems of Biological Classification." J. Roy. Statist. Soc., Ser. B, 10, 159-193 (1948).
- 29 Anderson, E. and Abbe, E.C. "A Quantitative Comparison of Specific and Generic Differences in the Betulaceae." J. Arnold Arboretum, 15, 43-49 (1934).
- 30 Anderson, E. and Whitaker, T.W. "Speciation in *Uvularia*." J. Arnold Arboretum, 15, 28-42 (1934).
- 31 Zubin, J. "A Technique for Measuring Likemindedness." J. abnorm soc. Psychol., 33, 508-516 (1938).
- 32 Thorndike, R.L. "Who Belongs in a Family?" Psychometrika, 18, 267-276 (1953).
- 33 Sneath, P.H.A. "Some Thoughts on Bacterial Classification." J. Gen. Microbiol., 17, 184-200 (1957).
- 34 Sneath, P.H.A. "The Application of Computers to Taxonomy." J. Gen. Microbiol., 17, 201-226 (1957).
- 35 Michener, C.D. and Sokal, R.R. "A Quantitative Approach to a Problem in Classification." Evolution, 11, 130-162 (1957).
- 36 Sokal R.R. and Michener, C.D. "A Statistical Method for Evaluating Systematic Relationships." Univ. Kansas Sci. Bull., 38, 1409-1438 (1958).
- 37 Johnson, L.A.S. "Rainbows End: The Quest for an Optimal Taxonomy." Systematic Zoology, 19, 203-239 (1970).
- 38 Blackwelder, R.E. "A Critique of Numerical Taxonomy." Systematic Zoology, 16, 64-72 (1967).
- 39 Sneath, P.H.A. "Recent Trends in Numerical Taxonomy." Taxon, 18, 14-20 (1969).

- 40 Williams, W.T. and Dale, M.B. "Fundamental Problems in Numerical Taxonomy," in: *Advances in Botanical Research II*, 35-68 ed. R.D. Preston, Academic Press, New York, London (1965).
- 41 Sokal, R.R., Camin, J.H., Rohlf, F.J. and Sneath, P.H.A. "Numerical Taxonomy: some points of view." *Systematic Zoology*, 14, 237-243 (1965).
- 42 Lambert, J.M. and Williams, W.T. "Multivariate Methods in Plant Ecology, IV. Nodal Analysis." *J. Ecol.*, 50, 775-802 (1962).
- 43 Tyron, R.C. and Bailey, D.E. "Cluster Analysis." McGraw-Hill Book Company, New York (1970).
- 44 Good, I.J. "Categorisation of Classification in Mathematics and Computer Science in Biology and Medicine. H.M.S.O. (1965).
- 45 Degerman, R. "Multidimensional Analysis of Complex Structure Mixtures of Class and Quantitative Variation." *Psychometrika*, 35, 475-491 (1970).
- 46 Sneath, P.H.A. and Sokal, R.R. "Numerical Taxonomy." W.H. Freeman and Company, San Francisco (1973).
- 47 Kowalski, B.R. and Bender, C.F. "Pattern Recognition. II. Linear and Nonlinear Methods for Displaying Chemical Data." *J. Amer. Chem. Soc.*, 95 (3) 686-693 (1973).
- 48 Cattell, R.B. "Factor Analysis." Harper, New York (1952). "The Three Basic Factor-Analytic Research Designs - their Interrelations and Derivatives." *Psychol. Bull.*, 49, 499-520 (1952).
- 49 Gower, J.C. "Some Distance Properties of Latent Root and Vector Methods used in Multivariate Analysis." *Biometrika*, 53, 325-338 (1966).
- 50 Williams, W.T. and Lance, G.N. "Logic of Computer-Based Intrinsic Classifications." *Nature*, 207, 159-161 (1965).
- 51 Pilowsky, I., Levine, S. and Boulton, D.M. "The Classification of Depression by Numerical Taxonomy." *Br. J. Psychiat.*, 115, 937-945 (1969).
- 52 Paykel, E.S. "Classification of Depressed Patients: A Cluster Analysis Derived Group." *Br. J. Psychiat.*, 118, 275 (1971).

- 53 Paykel, E.S. "Depressive Typology and Response to Amitriptyline." Br. J. Psychiat., 120, 147-156 (1972)
- 54 Ting, K.H., Lee, R.C.T., Milne, G.W.A., Shapiro, M. and Guarino, A.M. "Applications of Artificial Intelligence: Relationships between Mass Spectra and Pharmacological Activity of Drugs." Science, 180, 417-420 (1973).
- 55 Chu, K., Feldmann, R.J., Shapiro, M.B., Hazard, G.F., Jr., and Geran, R.I. "Pattern Recognition and Structure-Activity Relationship Studies. Computer-Assisted Prediction of Antitumor Activity in Structurally Diverse Drugs in an Experimental Mouse Brain Tumor System." J. Med. Chem., 18, 539-545 (1975).
- 56 Cain, A.J. "Logic and Memory in Linnaeus's System of Taxonomy." Proc. Linn. Soc. Lond., 169th Session, 144-163 (1958).
- 57 Cain, A.J. "The Evolution of Taxonomic Principles," in: Microbial Classification, 12th Symposium of the Society for General Microbiology, 1-13 eds. G.C. Ainsworth and P.H.A. Sneath, Cambridge University Press, Cambridge (1962).
- 58 Thompson, W.R. "The Philosophical Foundations of Systematics." Can. Entomol., 84, 1-16 (1952).
- 59 Jevons, W.S. "The Principles of Science: A Treatise on Logic and Scientific Method," 2nd edn., rev. Macmillan, London (1877).
- 60 Lindley, J. "A Natural System of Botany," 2nd edn. London (1836).
- 61 Bather, F.A. "Biological Classification: Past and Future." Quart. J. Geol. Soc. London., 83, Proc. Ixii-civ (1927).
- 62 Turrill, W.B. "Taxonomy and Phylogeny." Bot. Rev., 8, 247 (1942).
- 63 Lam, H.J. "Phylogenetic Symbols, Past and Present." Acta Biotheoretica, ser. A, 2, 154 (1936).
- 64 Lam, H.J. "Over de Eenheid der Bijzondere Plantkunde." Vokblad voor Biologen, 11, 201 (1938).
- 65 Gilmour, J.S.L. "Taxonomy and Philosophy," in: The New Systematics, 461-474 ed. J. Huxley, Clarendon Press, Oxford (1940).

- 66 Gilmour, J.S.L. "The Development of Taxonomic Theory since 1851." Nature, 168, 400-402 (1951).
- 67 Gilmour, J.S.L. "Taxonomy," in: Contemporary Botanical Thought, 27-45 eds. A.M. MacLeod and L.S., Copley, Oliver and Boyd, Edinburgh, and Quadrangle. Books, Chicago (1961).
- 68 Candolle, A.P.de. "Theorie elementaire de la botanique, ou exposition des principes de la classification naturelle et de l'art de decerne et d'etudier les vegetaux." Deterville, Paris (1813).
- 69 Beckner, M. "The Botanical Way of Thought." Columbia University Press, New York (1959).
- 70 Maccacaro, G.A. "La misura delle informazione contenuta nei criteri di classificazione." Ann. Microbiol. Enzimol., 8, 231-239 (1958).
- 71 Williams, W.T. and Lambert, J.M. "Multivariate Methods in Plant Ecology. I. Association Analysis in Plant Communities," J. Ecol. 47, 83-101 (1959).
- 72 Sammon, J.W., Jr. and Foley, D. "Considerations of Dimensionality versus Sample Size," in Proceedings of the IEEE Symposium on Adaptive Processes, IX, 2.1 - IX.2.7, University of Texas, Austin (1970).
- Kanal, L. and Chandrasekaran, B. "On Dimensionality and Sample Size in Statistical Pattern Recognition," in Proceedings of the National Electronics Conference, 2-7 (1968).
- 73 Foley, D.H. "Considerations of Sample and Feature Size." IEEE Transactions on Information Theory, IT-18(5), 618-626 (1972).
- 74 Jardine, N. and Sibson, R. "Mathematical Taxonomy." Wiley, London (1971).
- 75 Inglis, W.G. "The Purpose and Judgements of Biological Classifications." Systematic Zoology, 19, 240-250 (1970).
- 76 Colless, D.H. "'Phenetic', 'Phylogenetic' and 'Weighting'." Systematic Zoology, 20 (1), 73-76 (1971).
- 77 Goodall, D.W. "A New Similarity Index Based on Probability." Biometrics, 22, 882-907 (1966).

- 78 Smirnov, E.S. "On Exact Methods in Systematics." Systematic Zoology, 17, 1-13(1968).
- 79 Rogers, D.J. and Tanimoto, T.T. "A Computer Program for Classifying Plants." Science, 132, 1115-1118 (1960).
- 80 Kendrick, W.B. "Quantitative Characters in Computer Taxonomy," in: Phenetic and Phylogenetic Classification, 105-114 eds. V.H. Heywood and J. McNeil, Syst. Ass. Pub. (1964).
- 81 Kendrick, W.B. "Complexity and Dependence in Computer Taxonomy." Taxon, 14, 141-154 (1965).
- 82 Williams, W.T. "The Problem of Attribute-Weighting in Numerical Classification." Taxon, 18, 369-374 (1969).
- 83 Williams, W.T., Dale, M.B. and Macnaughton-Smith, P. "An Objective Method of Weighting in Similarity Analysis." Nature, 201, 426 (1963).
- 84 Rohlf, F.J. "Correlated Characters in Numerical Taxonomy." Systematic Zoology, 16, 109-126 (1967).
- 85 Power, D.M. "Statistical Analysis of Character Correlations in Brewer's Blackbirds." Systematic Zoology, 20, 186-203 (1971).
- 86 Boyce, A.J. "Mapping Diversity: A Comparative Study of Some Numerical Methods," in: Numerical Taxonomy. Proceedings of the Colloquium in Numerical Taxonomy, held in the University of St. Andrews, September 1968, 1-31 ed. A.J. Cole, Academic Press, London (1969).
- 87 Sokal, R.R. and Michener, C.D. "The Effects of Different Numerical Techniques on the Phenetic Classification of Bees of the Hoplitis Complex (Megachilidae)." Proc. Linn. Soc. Lond., 178, 59-74 (1967).
- 88 Williams, W.T., Lambert, J.M. and Lance, G.N. "Multivariate Methods in Plant Ecology. V. Similarity Analyses and Information-Analyses." J. Ecol., 54, 427-445 (1966).
- 89 Gower, J.C. "A General Coefficient of Similarity and Some of its Properties." Biometrics, 27, 857-872 (1971).
- 90 Sokal, R.R. "Distance as a Measure of Taxonomic Similarity." Systematic Zoology, 10, 70-79 (1961).

- 91 Penrose, L.S. "Distance, Size and Shape." Ann. Eugenics, 18, 337-343 (1954).
- 92 Bielecki, T. "Some Possibilities for Estimating Inter-Population Relationships on the Basis of Continuous Traits." Current Anthropol. 3, 3-8 (1962).
- 93 Wanke, A. "Metoda badan czestosci wystepowani zespolow cech czyli metoda stochastycznej korelacji wielorakiej." Przeglad antropologiczny, 19, 106-147 (1953).
- 94 Clark, P.J. "An Extension of the Coefficient of Divergence for use with Multiple Characters." Copeia, 2, 61-64 (1952).
- 95 Carmichael, J.W. and Sneath, P.H.A. "Taxometric Maps." Systematic Zoology, 18, 402-415 (1969).
- 96 Cain, A.J. and Harrison, G.A. "An Analysis of the Taxonomist's Judgement of Affinity." Proc. Zool. Soc. Lond., 131, 85-98 (1958).
- 97 Czekanowski, J. "Coefficient of Racial Likeness and 'Durschnittliche Differenz'". Anthrop. Anz., 9, 227-249 (1932).
- 98 Gower, J.C. "Multivariate Analysis and Multidimensional Goemetry." The Statistician, 17, 13-25 (1967).
- 99 Shepard, R.N. "The Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function. I." Psychometrika, 27, 125-139 (1962).
- 100 Shepard, R.N. "The Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function. II." Psychometrika, 27, 219-246 (1962).
- 101 Kruskal, J.B. "Multidimensional Scaling by Optimising Goodness of Fit to Nonmetric Hypothesis." Psychometrika, 29, 1-27 (1964).
- 102 Kruskal, J.B. "Nonmetric Multidimensional Scaling: A Numerical Method." Psychometrika, 29, 115-129 (1964).
- 103 Shannon, C.E. "A Mathematical Theory of Communications." Bell System Technical Journal, 27, 379-423, 623-656 (1948).
- 104 Estabrook, G.F. "An Information Theory Model for Character Analysis." Taxon, 16, 86-97 (1967).



- 105 Orloci, L. "Information Theory Models for Hierarchic and Non-Hierarchic Classifications," in: Numerical Taxonomy. Proceedings of the Colloquium in Numerical Taxonomy, held in the University of St Andrews, September 1968, 148-164 ed. A.J. Cole, Academic Press London (1969).
- 106 Hawksworth, F.G. Estabrook, G.F. and Rogers, D.J. "Application of an Information Theory Model for Character Analysis in the Genus *Arceuthobium* (Viscaceae)." Taxon, 17, 605-619 (1968).
- 107 Legendre, P. and Rogers, D.J. "Characters and Clustering in Taxonomy: A Synthesis of Two Taximetric Procedures." Taxon, 21, 567-606 (1972).
- 108 Fleiss, J.L. and Zubin, J. "On Methods and Theory of Clustering." Multivariate Behaviour Res., 4, 235-250 (1969).
- 109 Eades, D.C. "The Inappropriateness of the Correlation Coefficient as a Measure of Taxonomic Resemblance." Systematic Zoology, 14, 98-100 (1965).
- 110 Wishart, D. "A Generalised Approach to Cluster Analysis." Part of Ph.D. Thesis, University of St Andrews (1971).
- 111 Strauss, J.S., Bartko, J.J. and Carpenter, W.T. "The Use of Clustering Techniques for the Classification of Psychiatric Patients." Br. J. Psychiat., 122, 531-540 (1973).
- 112 Jardine, N. and Sibson, R. "A Model for Taxonomy." Math Biosci., 2, 465-482 (1968).
- 113 Parker-Rhodes, A.F. and Needham, R.M. "The Theory of Clumps." Cambridge Language Research Unit Working Papers, 126 (1960).
- 114 Needham, R.M. "The Theory of Clumps. II." Cambridge Language Research Unit Working Papers, 139 (1961).
- 115 Needham, R.M. and Spark Jones, K. "Keywords and Clumps." J. Documentation, 20, (1), 5-15 (1964).
- 116 Spark Jones, K. and Jackson, D.M. "Current Approaches to Classification and Clump Finding at the Cambridge Language Research Unit." The Computer Journal, 10, 29-37 (1967).
- 117 Needham, R.M. "Automatic Classification in Linguistics." The Statistician, 17, 45-54 (1967).

- 118 Needham, R.M. "Research on Information Retrieval, Classification and Clumping." Ph.D. Thesis, Cambridge, (1961).
- 119 Beale, E.M.L. "Cluster Analysis." Scientific Control Systems, London (1969).
- 120 Calinski, T. and Harabasz, J. "A Dendrite Method for Cluster Analysis." Unpublished Manuscript (1971).
- 121 Marriot, F.H.C. "Practical Problems in a Method of Cluster Analysis." Biometrics, 27, 501-514 (1971).
- 122 Wolfe, J.H. "A Monte Carlo Study of the Sampling Distribution of the Likelihood Ratio for Mixtures of Multinormal Distributions." Naval Personnel and Training Research Laboratory. Technical Bulletin, STB 72-2, San Diego, California (1971).
- 123 Jardine, N. and Sibson, R. "The Construction of Hierarchic and Non-Hierarchic Classifications." The Computer Journal, 11, 177-184 (1968).
- 124 Cole, A.J. and Wishart, D. "An Improved Algorithm for the Jardine-Sibson Method of Generating Overlapping Clusters." The Computer Journal, 13, 156-163 (1970).
- 125 Rohlf, F.J. "A New Approach to the Computation of Jardine and Sibson's Bk Clusters." MS (1973).
- 126 Lambert, J.M. and Williams, W.T. "Multivariate Methods in Plant Ecology. VI Comparison of Information-Analysis and Association Analysis." J. Ecol., 54, 635-664 (1966).
- 127 Macnaughton-Smith, P. "Some Statistical and Other Numerical Techniques for Classifying Individuals." Home Office Research Unit Report, No 6. H.M.S.O. London (1965).
- 128 Florek, K., Lukaszewicz, J., Perkal, J., Steinhaus, H. and Zubrzycki, S. "Sur la liaison et la division des points d'un ensemble fini." Colloquium Math., 2, 282-285 (1951).
- 129 Wishart, D. "Mode Analysis," in: Numerical Taxonomy. Proceedings of the Colloquium in Numerical Taxonomy, held in the University of St Andrews, September 1968, 282-308 ed. A.J. Cole, Academic Press, London (1969).
- 130 Wishart, D. "Numerical Classification Method for Deriving Natural Classes." Nature, 221, 97-98 (1969).

- 131 Carmichael, J.W., George, J.A. and Julius, R.S. "Finding Natural Clusters." Systematic Zoology, 17, 144-150 (1968).
- 132 Gitman, I. and Levine, M.D. "An Algorithm for Detecting Unimodal Fuzzy Sets and its Application as a Clustering Technique." IEEE Trans. Comput., C19, 583-593 (1970).
- 133 Cattell, R.B. and Coulter, M.A. "Principles of Behavioural Taxonomy and the Mathematical Basis of the Taxonome Computer Program." Br. J. Math. Statist. Psychol., 19, 237-269 (1966).
- 134 Day, N.E. "Estimating the Components of a Mixture of Normal Distributions." Biometrika, 56, 463-474 (1969).
- 135 Sneath, P.H.A. "Evaluation of Clustering Methods" in: Numerical Taxonomy. Proceedings of the Colloquium in Numerical Taxonomy, held in the University of St Andrews, September 1968, 257-271 ed. A.J. Cole, Academic Press, London (1969).
- 136 Wolfe, J.H. "Pattern Clustering by Multivariate Mixture Analysis." Multiv. Behav. Res., 5, 329-350 (1970).
- 137 Jardine, C.J., Jardine, N. and Sibson, R. "The Structure and Construction of Taxonomic Hierarchies." Math. Biosci., 1, 173-179 (1967).
- 138 Williams, W.T., Lance, G.N., Dale, M.B. and Clifford, H.T. "Controversy Concerning the Criteria for Taxonomic Strategies." The Computer Journal, 14, 162-165 (1971).
- 139 Forgey, E.W. "Cluster Analysis of Multivariate Data: Efficiency versus Interpretability of Classification." Biometrics, 21, 768-769 (1965).
- 140 Sibson, R. Discussion on R.M. Cormack's Paper - A Review of Classification. Jl. R. Statist. Soc., Series A, 134, (3), 321-367 (1971).
- 141 Cunningham, K.M. and Ogilvie, J.C. "Evaluation of Hierarchical Grouping Techniques: A Preliminary Study." The Computer Journal, 15, 209-213 (1972).
- 142 Farris, J.S. "On the Cophenetic Correlation Coefficient." Systematic Zoology, 18, 279-285 (1969).
- 143 Rohlf, R.J. "Adaptive Hierarchical Clustering Schemes." Systematic Zoology, 19, 58-82 (1970).

- 144 Johnson, S.C. "Hierarchical Clustering Schemes." Psychometrika, 32, 241-254 (1967).
- 145 Hartigan, J.A. "Representation of Similarity Matrices by Trees." J. Amer. Stat. Ass., 62, 1140-1158 (1967).
- 146 Wallace, C.S. and Boulton, D.M. "An Information Measure for Classification." The Computer Journal, 11, 185-194 (1968).
- 147 Sokal, R.R. and Rohlf, F.J. "The Comparison of Dendrograms by Objective Methods." Taxon, 11, 33-40 (1962).
- 148 Bonner, R.E. "On Some Clustering Techniques." IBM J. Res. Dev., 8, 22-32 (1964).
- 149 Kier, L.B. "Molecular Orbital Theory in Drug Research." Medicinal Chemistry Series, 10, Academic Press, New York (1971).
- 150 Cammarata, A. and Martin, A.N. in: Medicinal Chemistry 3rd. Ed., 1, 118-163 ed. A Burger, Wiley, New York (1970).
- 151 Burger, A. "The Interpretation of Structure-Activity Relationships," in: Fundamental Concepts in Drug-Receptor Interactions, 1-13 eds. J.F. Danielli, J.F. Moran and D.J. Triggle, Academic Press, New York (1970).
- 152 Mautner, H.G. Annu. Rep. Med. Chem., 230 (1969).
- 153 Chu, K. "Application of Artificial Intelligence to Chemistry." Analytical Chemistry, 46, (9), 1181-1187 (1974).
- 154 Cramer, R.D.III, Redl, G. and Berkoff, C.E. "Substructural Analysis. A Novel Approach to the Problem of Drug Design." J. Med. Chem., 17 (5), 533-535 (1974).
- 155 Harrison, P.J. "A Method of Cluster Analysis and Some Applications." J. Applied Statistics, 17, 226-236 (1968).
- 156 Sagers, D.T. "The Application of the Computer to a Pesticide Screening Programme." Pesticide Science, 5, 341-352 (1974).
- 157 Isenhour, T.L. and Jurs, P.C. "Learning Machines," in: Computer Fundamentals for Chemists, 1, 285-331 eds. J.S. Mattson, H.B. Mark, Jr. and H.C. Macdonald, Jr., M. Dekker inc., New York (1973).

- 158 Johnson, J.E. and Blair, E.H. "Cost, Time and Pesticide Safety." Chem. Technol., 666-669, Nov 1972.
- 159 Arnett, E.M. "Computer-Based Chemical Information Services." Science, 170, 1370-1376 (1970).
- 160 Ash, J.E. and Hyde, E. "Chemical Information Systems." Chichester, Ellis Horwood (1975).
- 161 Richet, M.C., C. R. Soc. Biol. (Paris), 45, 775 (1893).
- 162 Meyer, H. Arch. Expt. Pathol. Pharmacol., 42, 109 (1899).
- 163 Overton, E. Z. Phys. Chem., 22, 189 (1897).
- 164 Cavallito, C.J. "Approaches to Drug Design," in: Medicinal Chemistry 3rd. edn., I, 233-245 ed. A. Burger, Wiley, New York (1970).
- 165 Seydel, J.K. "Physicochemical Approaches to the Rational Development of New Drugs," in: Drug Design I, 343-380 ed. E.J. Ariens, Medicinal Chemistry Series, 11, Academic Press, New York (1971).
- 166 Albert, A. "Relations between Molecular Structure and Biological Activity: Stages in the Evolution of Current Concepts." Annu. Rev. Pharmacol., 11, 13-36 (1971).
- 167 Hammett, L.P. "The Effect of Structure upon the Reactions of Organic Compounds." Benzene Derivatives. J. Amer. Chem. Soc., 59, 96-103 (1937).
- 168 Hansch, C. and Yoshimoto, M. "Structure-Activity Relationships in Immunochemistry. 2. Inhibition of Complement by Benzamidines." J. Med. Chem., 17 (11), 1160-1167 (1974).
- 169 Hansch, C. "A Quantitative Approach to Biochemical Structure-Activity Relationships." Accounts Chem. Res., 2, 232-239 (1969).
- 170 Hansch, C. "Physicochemical Parameters in Drug Design." Annu. Rep. Med. Chem., 348-357, Academic Press, New York (1967).
- 171 Clayton, J.M., Millner, O.E., Jr. and Purcell, W.P. "Physicochemical Parameters in Drug Design." Annu. Rep. Med. Chem., 285-295, Academic Press, New York (1970).

- 172 Martin, Y.C., Holland, J.B., Jarboe, C.H. and Plotnikoff, N. "Discriminant Analysis of the Relationship between Physical Properties and the Inhibition of Monoamine Oxidase by Amino-tetralins and Aminoindans." J. Med. Chem., 17 (4), 409-413 (1974).
- 173 Fujita, T., Iwasa, J. and Hansch, C. "A New Substituent Constant,  $\Pi$ , Derived from Partition Coefficients." J. Amer. Chem. Soc., 86, 5175-5180 (1964).
- 174 Currie, D.J., Lough, C.E., Silver, R.F. and Holmes, H.L. "Partition Coefficients of Some Conjugated Heteroenoid Compounds and 1,4-Naphthoquinones." Can. J. Chem., 44, 1035-1043 (1966).
- 175 Iwasa, J., Fujita, T. and Hansch, C. "Substituent Constants for Aliphatic Functions Obtained from Partition Coefficients." J. Med. Chem., 8, 150-153 (1965).
- 176 Hansch, C. and Anderson, S.M. "The Effect of Intramolecular Hydrophobic Bonding on Partition Coefficients." J. Org. Chem., 32, 2583-2586 (1967).
- 177 Scholtan, W. "Die Bindung der Antibiotica an die Eiweisskorper des Serums." Arzneimittel-Forsch, 13, 347-360 (1963).
- 178 Purcell, W.P., Singer, J.A., Sundaram, K. and Parks, G.L. "Quantitative Structure-Activity Relationships and Molecular Orbitals in Medicinal Chemistry," in: Medicinal Chemistry 3rd. edn., I, 164-192 ed. A. Burger, Wiley, New York (1970).
- 179 Hammett, L.P. "Physical Organic Chemistry." McGraw-Hill, New York (1940).
- 180 Hermann, R.B., Culp, H.W., McMahon, R.E. and Marsh, M.M. "Structure-Activity Relationships among Substrates for a Rabbit Kidney Reductase. Quantum Chemical Calculation of Substituent Parameters." J. Med. Chem., 12, 749-754 (1969).
- 181 Hansch, C., Unger, S.H. and Forsythe, A.B. "Strategy in Drug Design. Cluster Analysis as an Aid in the Selection of Substituents." J. Med. Chem., 16 (11), 1217-1222 (1973).
- 182 Mathews, R.J. "A Comment on the Structure-Activity Correlations obtained using Pattern Recognition Methods." J. Amer. Chem. Soc., 97 (4), 935-936 (1975).

- 183 Kowalski, B.R. and Bender, C.F. "Computer used to Evaluate Anti-cancer Drugs." Chem. Eng. News, 52 (7), 19 (1974).
- 184 Sammon, J.W., Jr. "A Non-linear Mapping for Data Structure Analysis." IEEE Trans. Comput., C18, 401-409 (1969).
- 185 Bruice, T.C., Kharasch, N. and Winzler, R.J. "A Correlation of Thyroxine-Like Activity and Chemical Structure." Archives of Biochemistry and Biophysics, 62, 305-317 (1956).
- 186 Fujita, T and Ban, T. "Structure-Activity Study of Phenethylamines as Substrates of Biosynthetic Enzymes of Sympathetic Transmitters." J. Med. Chem., 4, 148-152 (1971).
- 187 Purcell, W.P. "Cholinesterase Inhibitory Prognoses of Thirty Six Alkyl Substituted 3-Carbamoylpiperidines." Biochim. Biophys. Acta 105, 201-204 (1965).
- 188 Purcell, W.P. and Clayton, J.M. "Application of Regression Analyses to Antitumor Activities of Various Acetylenic Carbamates." J. Med. Chem., 11 (2), 199-203 (1968).
- 189 Sokal, R.R. and Rohlf, F.J. "Introduction to Biostatistics." W.H. Freeman and Company, San Francisco (1969).
- 190 Hancock, C.K. "Some Misconceptions of Regression Analysis in Physical Organic Chemistry." J. Chem. Education, 42 (11), 608-609 (1965).
- 191 Goodall, D.W. "Hypothesis-testing in Classification." Nature, 211, 329-330 (1966).
- 192 Wishart, D. "FORTRAN II Programs for 8 Methods of Cluster Analysis (CLUSTAN I)." Kansas Geol. Surv. Computer Contrib., No. 38 (1969).
- 193 Wishart, D. "Some Problems in the Theory and Application of the Methods of Numerical Taxonomy." Ph.D. Thesis, University of St Andrews (1970).
- 194 Rohlf, F.J., Kishpaugh, J. and Kirk, D., "NT-SYS. Numerical Taxonomy System of Multivariate Statistical Programs." Tech. Rep. State University of New York (1971).
- 195 Lance, G.N. and Williams, W.T. "Computer Programs for Hierarchical Polythetic Classification ('similarity analyses')." The Computer Journal, 9, 60-64 (1966).

- 196 Ross, G.J.S. "Algorithms AS13 (Minimum Spanning Tree), AS14 (Printing Minimum Spanning Tree) and AS15 (Single-Linkage Cluster Analysis)." J. Applied Statistics, 18, 103-110 (1969).
- 197 Gower, J.C. and Ross, G.J.S. "Minimum Spanning Trees and Single Linkage Cluster Analysis." J. Applied Statistics, 18, 54-64 (1969).
- 198 Rose, M.J. "Classification of a Set of Elements." The Computer Journal, 7, 208-211 (1964).
- 199 Ceska, A. "Application of Association Coefficients for Estimating the Mean Similarity between Sets of Vegetational Relevés." Folia Geobot. Phytotaxonom., 3, 57-64 (1968).
- 200 Ross, G.J.S. "Classification Techniques for Large Sets of Data." in: Numerical Taxonomy. Proceedings of the Colloquium in Numerical Taxonomy, held in the University of St Andrews, September 1968, 282-308 ed. A.J. Cole, Academic Press, London (1969).
- 201 Lockhart, W.R. and Hartman, P.A. "Formation of Monothetic Groups in Quantitative Bacterial Taxonomy." J. Bacteriol., 85, 68-77 (1963).
- 202 Crawford, R.M.M. and Wishart, D. "A Rapid Multivariate Method for the Detection and Classification of Groups of Ecologically Related Species." J. Ecol., 55, 505-524 (1967).
- 203 Crawford, R.M.M. and Wishart, D. "A Rapid Classification and Ordination Method and its Application to Vegetation Mapping." J. Ecol., 56, 385-404 (1968).
- 204 Kaminuma, T., Takekawa, T. and Watanabe, S. "Reduction of Clustering Problem to Pattern Recognition." Pattern Recognition, 1, 195-205, (1969).
- 205 Switzer, P. "Numerical Classification" in: Geostatistics, a Colloquium, 31-43 ed. D.F. Merriam, Plenum Press, New York (1970).
- 206 Sibson, R. SLINK: "An Optimally Efficient Algorithm for the Single-Link Cluster Method." The Computer Journal, 16, 30-34 (1973).
- 207 Adamson, G.W., Bush, J.A., Mclure, A.H. and Lynch, M.F. An "Evaluation of a Substructure Search System based on Bond-Centred Fragments." J. Chem. Documentation, 14 (1), 44-48 (1974).



- 208 Adamson, G.W., Lynch, M.F. and Town, W.G. "Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-Based File, II, Atom-centred Fragments." J. Chem. Soc., C, 3702-3706 (1971).
- 209 Agin, D., Hersh, L. and Holtzman, D. "The Action of Anaesthetics on Excitable Membranes: A Quantum-Chemical Analysis." Proc. N.A.S., 53, 952-958 (1965).
- 210 Bird, A.E. and Marshall, A.C. "Correlation of Serum of Penicillins with Partition Coefficients." Biochem. Pharmacol., 16, 2275-2289 (1967).
- 211 Meister, A. "The Biochemistry of the Amino Acids." Academic Press, New York (1957).
- 212 Crowe, J.E., Lynch, M.F. and Town, W.G. "Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-Based File, I, Non-cyclic Fragments." J. Chem. Soc., C, 990-996 (1970).
- 213 Adamson, G.W., Creasey, S.E. and Lynch, M.F. "Analysis of Structural Characteristics of Chemical Compounds in the Common Data Base." J. Chem. Documentation, 13 (3), 158-162 (1973).
- 214 Nys, G.G. and Rekker, R.F. "Statistical Analysis of a Series of Partition Coefficients with Special Reference to the Predictability of Folding of Drug Molecules." Chimie Therapeutique, 5, 521-535 (1973).
- 215 Hansch, C. and Anderson, S.M. "The Structure-Activity Relationships in Barbiturates and Its Similarity to That in Other Narcotics." J. Med. Chem., 10, 745-753 (1967).
- 216 Muir, R.M., Fujita, T. and Hansch, C. "Structure-Activity Relationship in the Auxin Activity of Mono-Substituted Phenylacetic Acids." Plant Physiol., 42, 1519-1526 (1967).
- 217 Singer, J.A. and Purcell, W.P. "Relationships among Current Quantitative Structure-Activity Models." J. Med. Chem., 10, 1000-1002 (1967).
- 218 Hansch, C., Steward, A.R., Anderson, S.M. and Bentley, D. "The Parabolic Dependence of Drug Action upon Lipophilic Character as Revealed by a Study of Hypnotics." J. Med. Chem., 11, 1-11 (1967).
- 219 Statistical Analysis Mark II Applications Package, International Computers Limited Technical Publication 4301, London (1971).

- 220 Zaagsma, J., and Nauta, W.Th." -Adrenoceptor Studies. I. In Vitro  $\beta$ -Adrenergic Blocking, Antiarrhythmic and Local Anaesthetic Activities of a New Series of Aromatic Bis (2-Hydroxy-3-Isopropyl-aminopropyl)Ethers." J. Med. Chem., 17, 507-513 (1974).
- 221 Adamson, G.W. and Bawden, D. "A Method of Structure-Activity Correlation using Wiswesser Line Notation." J. Chem. Inf. Comput. Sci., 15, 215-220 (1975).
- 222 Adamson, G.W. and Bawden, D. "An Empirical Method of Structure-Activity Correlation for Polysubstituted Cyclic Compounds using Wiswesser Line Notation." J. Chem. Inf. Comput. Sci., 1976, in the press.
- 223 Bawden, D. Doctoral Thesis, University of Sheffield, in preparation
- 224 Leiter, D.P. and Leighner, L.H. A Statistical Analysis of the Structure Registry at Chemical Abstracts Service, presented at the ACS Meeting, Chicago (1967).
- 225 Unpublished results of work carried out at the Chemotherapeutic Research Centre, Beecham Pharmaceuticals, Betchworth, Surrey, England.

pI	Amino acid	Ala	Arg	Asp	Asp (NH <sub>2</sub> )	Cys	Glu	Glu (NH <sub>2</sub> )	Gly	His	Ileu	Leu	Lys	Met	Phen	Pro	Ser	Thr	Try	Tyr	Val
6.00	Alanine		0.503	0.621	0.621	0.734	0.577	0.577	0.693	0.537	0.784	0.784	0.577	0.621	0.503	0.331	0.734	0.844	0.416	0.471	0.844
10.76	Arginine		0.452	0.578	0.434	0.524	0.645	0.586	0.359	0.452	0.452	0.766	0.452	0.318	0.114	0.434	0.377	0.213	0.281	0.377	
2.77	Aspartic acid			0.722	0.552	0.935	0.668	0.530	0.491	0.583	0.583	0.535	0.583	0.452	0.203	0.705	0.639	0.354	0.539	0.494	
5.41	Asparagine				0.552	0.668	0.935	0.530	0.491	0.583	0.583	0.668	0.583	0.452	0.203	0.552	0.494	0.354	0.417	0.494	
5.07	Cysteine					0.508	0.508	0.629	0.469	0.552	0.552	0.508	0.705	0.434	0.282	0.662	0.602	0.347	0.403	0.602	
3.22	Glutamic acid						0.743	0.491	0.442	0.535	0.535	0.614	0.535	0.403	0.171	0.655	0.590	0.302	0.485	0.450	
5.65	Glutamine							0.491	0.442	0.535	0.535	0.743	0.535	0.403	0.171	0.508	0.450	0.302	0.367	0.450	
5.97	Glycine								0.457	0.530	0.530	0.661	0.530	0.426	0.390	0.629	0.575	0.350	0.399	0.575	
7.59	Histidine									0.491	0.491	0.442	0.491	0.359	0.277	0.469	0.412	0.475	0.322	0.412	
6.02	Isoleucine										1.000	0.535	0.583	0.452	0.203	0.552	0.639	0.354	0.417	0.930	
5.98	Leucine											0.535	0.583	0.452	0.203	0.552	0.639	0.354	0.417	0.930	
9.74	Lysine												0.535	0.403	0.171	0.508	0.450	0.302	0.367	0.450	
5.74	Methionine													0.452	0.203	0.552	0.494	0.354	0.417	0.494	
5.48	Phenylalanine														0.114	0.434	0.377	0.640	0.835	0.377	
6.30	Proline															0.282	0.240	0.164	0.088	0.240	
5.68	Scrine																0.763	0.347	0.538	0.602	
6.16	Threonine																	0.288	0.473	0.696	
5.89	Tryptophan																		0.589	0.288	
5.66	Tyrosine																				0.345
5.96	Valine																				

Table 1 pI and  $\phi$  values for 20 amino acids. The  $\phi$  values were calculated using structure representation (ii), (augmented atoms).

pI	Amino acid	Ala	Arg	Asp	Asp (NH <sub>2</sub> )	Cys	Glu	Glu (NH <sub>2</sub> )	Gly	His	Ileu	Leu	Lys	Met	Phen	Pro	Ser	Thr	Try	Tyr	Val
6.00	Alanine		0.503	0.621	0.621	0.734	0.577	0.577	0.693	0.537	0.784	0.784	0.577	0.621	0.503	0.331	0.734	0.844	0.416	0.471	0.844
10.76	Arginine		0.452	0.578	0.434	0.524	0.645	0.586	0.359	0.452	0.452	0.766	0.452	0.318	0.114	0.434	0.377	0.213	0.281	0.377	
2.77	Aspartic acid			0.722	0.552	0.935	0.668	0.530	0.491	0.583	0.583	0.535	0.583	0.452	0.203	0.705	0.639	0.354	0.539	0.494	
5.41	Asparagine				0.552	0.668	0.935	0.530	0.491	0.583	0.583	0.668	0.583	0.452	0.203	0.552	0.494	0.354	0.417	0.494	
5.07	Cysteine					0.508	0.508	0.629	0.469	0.552	0.552	0.508	0.705	0.434	0.282	0.662	0.602	0.347	0.403	0.602	
3.22	Glutamic acid						0.743	0.491	0.442	0.535	0.535	0.614	0.535	0.403	0.171	0.655	0.590	0.302	0.485	0.450	
5.65	Glutamine							0.491	0.442	0.535	0.535	0.743	0.535	0.403	0.171	0.508	0.450	0.302	0.367	0.450	
5.97	Glycine								0.457	0.530	0.530	0.661	0.530	0.426	0.390	0.629	0.575	0.350	0.399	0.575	
7.59	Histidine									0.491	0.491	0.442	0.491	0.359	0.277	0.469	0.412	0.475	0.322	0.412	
6.02	Isoleucine											1.000	0.535	0.583	0.452	0.203	0.552	0.639	0.354	0.417	0.930
5.98	Leucine												0.535	0.583	0.452	0.203	0.552	0.639	0.354	0.417	0.930
9.74	Lysine													0.535	0.403	0.171	0.508	0.450	0.302	0.367	0.450
5.74	Methionine														0.452	0.203	0.552	0.494	0.354	0.417	0.494
5.48	Phenylalanine														0.114	0.434	0.377	0.640	0.835	0.377	
6.30	Proline																0.282	0.240	0.164	0.088	0.240
5.68	Serine																	0.763	0.347	0.538	0.602
6.16	Threonine																		0.288	0.473	0.696
5.89	Tryptophan																			0.589	0.288
5.66	Tyrosine																				0.345
5.96	Valine																				

Table 1 pI and  $\phi$  values for 20 amino acids. The  $\phi$  values were calculated using structure representation (ii), (augmented atoms).

SC or DC type	Structural representation		
	(i)	(ii)	(iii)
Dice SC	0.81	0.39	0.42
	-	0.43	0.48
Sneath DC	0.81	0.74	0.76
	-	0.50	0.46
∅	0.81	0.39	0.42
	-	0.43	0.48

Table 2 Mean differences between observed and 'predicted pI values for 20 naturally occurring amino acids using three types of SC and DC and three structural representations in terms of augmented atoms. The upper values in the cells were calculated from the average pI value of the cluster which an acid joined and the lower entry from the acid(s) with which the acid with 'unknown' pI value has the highest SC or lowest DC.

Structural representation  SC or DC type	(i)	(ii)
	Dice SC	1.53
1.46		0.99
Sneath DC	1.32	1.30
	1.27	0.84
∅	1.56	1.27
	1.51	1.07

Table 3 Mean differences between observed and 'predicted' log (MBC) values for 39 local anaesthetics using three types of SC and DC and three structural representations in terms of augmented atoms. For an explanation of the different cell entries see Table 2.

Measure of Association	Predictions based on highest SC or lowest DC			Predictions based on classification		
	$\frac{\sum_{i=1}^n  x_i - \hat{x}_i }{\sum_{i=1}^n  x_i - \bar{x} }$	$\frac{\sum_{i=1}^n (x_i - \hat{x}_i)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$	$\frac{\sum_{i=1}^n  x_i - \hat{x}_i }{n}$	$\frac{\sum_{i=1}^n  x_i - \hat{x}_i }{\sum_{i=1}^n  x_i - \bar{x} }$	$\frac{\sum_{i=1}^n (x_i - \hat{x}_i)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$	$\frac{\sum_{i=1}^n  x_i - \hat{x}_i }{n}$
1	0.603	0.527	0.994	0.707	0.543	1.167
2	0.650	0.662	1.071	0.772	0.664	1.273
3	0.511	0.428	0.843	0.788	0.819	1.300
4	0.477	0.343	0.786	0.653	0.659	1.076
4(a)	0.648	0.611	1.069	0.867	0.859	1.429
5	0.882	0.928	1.454	0.958	0.950	1.579
6	1.147	1.618	1.891	1.196	1.420	1.973
7	0.602	0.516	1.000	0.732	0.679	1.207

Table 4 Log (MBC) estimations for a group of 39 local anaesthetics, based on a number of different measures of association and the classifications obtained using these.  $x_i$  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.

Fragment type	Predictions based on highest SC			Predictions based on classification		
	$\frac{n}{\sum_{i=1}^n  x_i - \hat{x}_i }$	$\frac{n}{\sum_{i=1}^n (x_i - \hat{x}_i)^2}$	$\frac{n}{\sum_{i=1}^n  x_i - \hat{x}_i }$	$\frac{n}{\sum_{i=1}^n  x_i - \hat{x}_i }$	$\frac{n}{\sum_{i=1}^n (x_i - \hat{x}_i)^2}$	$\frac{n}{\sum_{i=1}^n  x_i - \hat{x}_i }$
	$\frac{n}{\sum_{i=1}^n  x_i - \bar{x} }$	$\frac{n}{\sum_{i=1}^n (x_i - \bar{x})^2}$	n	$\frac{n}{\sum_{i=1}^n  x_i - \bar{x} }$	$\frac{n}{\sum_{i=1}^n (x_i - \bar{x})^2}$	n
Atoms	0.774	0.533	0.796	0.807	0.528	0.830
Augmented atoms	0.421	0.125	0.428	0.384	0.111	0.395
Simple pairs (SP)	0.879	0.704	0.904	1.025	1.010	1.054
Augmented pairs	0.879	0.774	0.904	0.768	0.849	0.790
Bonded pairs (BP)	0.803	0.613	0.826	0.838	0.899	0.862
Octuplets (OC)	0.684	0.446	0.704	0.404	0.111	0.416
SP + BP	0.663	0.525	0.682	0.770	0.864	0.792
SP + BP + OC	0.635	0.460	0.653	0.833	1.074	0.857

Table 5 PI estimations for 20 naturally occurring amino acids, based on Dice's SC and the classifications obtained with this coefficient, using a variety of different fragment definitions. Quantities  $x_i$ ,  $\hat{x}_i$ ,  $\bar{x}$  and  $n$  are defined in Table 4.



Fragment type	Predictions based on highest SC			Predictions based on classification		
	$\frac{\sum_{i=1}^n  x_i - \hat{x}_i }{\sum_{i=1}^n  x_i - \bar{x} }$	$\frac{\sum_{i=1}^n (x_i - \hat{x}_i)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$	$\frac{\sum_{i=1}^n  x_i - \hat{x}_i }{n}$	$\frac{\sum_{i=1}^n  x_i - \hat{x}_i }{\sum_{i=1}^n  x_i - \bar{x}_i }$	$\frac{\sum_{i=1}^n (x_i - \hat{x}_i)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$	$\frac{\sum_{i=1}^n  x_i - \hat{x}_i }{n}$
Atoms	0.384	0.133	0.633	0.449	0.214	0.740
Augmented atoms	0.603	0.527	0.994	0.707	0.543	1.167
Simple pairs (SP)	0.418	0.301	0.690	0.546	0.339	0.899
Augmented pairs	0.686	0.726	1.130	0.639	0.456	1.053
Bonded pairs (BP)	0.797	0.899	1.314	0.767	0.579	1.264
Octuplets (OC)	0.786	0.824	1.296	0.796	0.667	1.312
SP + BP	0.529	0.457	0.873	0.651	0.459	1.073
SP + BP + OC	0.564	0.492	0.930	0.725	0.586	1.196

Table 6 Log (MBC) estimations for 39 local anaesthetics under the conditions specified in Table 5. Quantities  $x_i$ ,  $\hat{x}_i$ ,  $\bar{x}$  and  $n$  are defined in Table 4.

Fragment type	Predictions based on highest SC			Predictions based on classification		
	$\frac{n \sum_{i=1}  x_i - \hat{x}_i }{n \sum_{i=1}  x_i - \bar{x} }$	$\frac{n \sum_{i=1} (x_i - \hat{x}_i)^2}{n \sum_{i=1} (x_i - \bar{x})^2}$	$\frac{n \sum_{i=1}  x_i - \hat{x}_i }{n}$	$\frac{n \sum_{i=1}  x_i - \hat{x}_i }{n \sum_{i=1}  x_i - \bar{x} }$	$\frac{n \sum_{i=1} (x_i - \hat{x}_i)^2}{n \sum_{i=1} (x_i - \bar{x})^2}$	$\frac{n \sum_{i=1}  x_i - \hat{x}_i }{n}$
Atoms	0.614	0.393	0.335	0.709	0.574	0.387
Augmented atoms	0.683	0.475	0.373	0.694	0.495	0.379
Simple pairs (SP)	0.591	0.343	0.323	0.719	0.572	0.392
Augmented pairs	0.693	0.473	0.378	0.821	0.631	0.448
Bonded pairs (BP)	0.670	0.434	0.366	0.783	0.574	0.427
Octuplets (OC)	0.677	0.506	0.370	0.707	0.540	0.386
SP + BP	0.641	0.399	0.350	0.801	0.646	0.437
SP + BP + OC	0.645	0.404	0.352	0.694	0.508	0.379

Table 7 Log (B/F) estimations for 79 penicillins, under the conditions specified in Table 5. Quantities  $x_i$ ,  $\hat{x}_i$ ,  $\bar{x}$  and  $n$  are defined in Table 4.

Fragment type used for classification	Predicted log(MBC)				
	Acetone	Isopropanol	Propanol	Urethane	Ethyl ether
Simple pairs	1.66	2.40	2.55	1.72	1.22
Augmented pairs	2.55	2.60	2.60	-0.52	-0.09
Bonded pairs	-0.20	-0.12	0.72	-0.65	-0.15
Octuplets	-0.12	-0.20	0.72	1.33	1.34
Observed log(MBC)	2.60	2.55	2.40	2.00	1.93

Table 8

Examples of the classification results in 39 local anaesthetics for the smaller acyclic structures, showing the general improvement in the level of 'prediction' as the fragment size decreases

Fragment type used for classification	Predicted log (MBC)				
	Eserine	Quinine	Dibucaine	Diethyl ether	Butanol
Atoms	-2.90	-4.20	-3.60	1.78	1.93
Simple pairs	-0.56	-0.65	-2.18	1.22	0.45
Augmented pairs	-0.55	-0.42	-2.90	-0.10	0.45
Bonded pairs	-0.09	-0.58	-2.86	-0.15	0.45
Octuplets	-0.65	-0.58	-0.12	1.34	0.45
Observed log(MBC)	-3.66	-3.60	-4.20	1.93	1.78

Table 9

Examples of the classification results in 39 local anaesthetics, showing some of the improvements obtained with atom descriptions

Independent variables		Significance level	Variables included in regression	Degrees of freedom	Multiple correlation coefficient	Residual error	F statistic
Type	No						
Atoms	4	99%	3 + const	35	0.979	0.433	266.12 (35,3)
		10%	2 + const	36	0.979	0.429	410.58 (36,2)
Simple pairs	16	99%	14 + const	24	0.995	0.261	169.71 (24,14)
		10%	11 + const	27	0.994	0.258	202.08 (27,11)
Simple pairs + Squared terms	24	99%	19 + const	19	0.995	0.284	99.00 (19,19)
		10%	9 + const	29	0.993	0.280	226.94 (29,9)
Simple pairs (Quadratic)	36	99%	27 + const	11	0.998	0.232	101.44 (11,27)
		10%	13 + const	25	0.996	0.240	238.46 (25,13)
Augmented pairs	36	99%	30 + const	8	0.999	0.228	133.07 (8,30)
		10%	18 + const	20	0.998	0.164	276.67 (20,18)
$\alpha$ I (Agin)	1	-	1 + const	37	0.993	0.240	260.51 (37,1)

Table 10 Summary of the empirical regression results in 39 local anaesthetics, including the semi-empirical result obtained by Agin et al<sup>209</sup> for the same structures.

Independent variables		Significance level	Variables included in regression	Degrees of freedom	Multiple correlation coefficient	Residual error	F statistic
Type	No						
Atoms	7	99%	6 + const	72	0.877	0.333	39.96 (72,6)
		10%	4	75	0.920	0.332	102.94 (75,4)
Simple pairs	14	99%	12 + const	66	0.912	0.297	27.23 (66,12)
		10%	7	72	0.932	0.315	68.19 (72,7)
Augmented pairs	33	99%	28	51	0.979	0.208	41.55 (51,28)
		10%	14	65	0.976	0.199	94.16 (65,14)
Bonded pairs	51	99%	39	40	0.982	0.221	27.47 (40,39)
		10%	20	59	0.975	0.212	57.26 (59,20)
Octuplets	71	99%	52	27	0.987	0.184	19.45 (27,52)
		10%	25	54	0.986	0.164	74.97 (54,25)
Augmented atoms	44	99%	29	50	0.976	0.225	34.97 (50,29)
		10%	14	65	0.972	0.215	79.76 (65,14)
ΣII (Bird & Marshall)	1	-	1 + const	77	0.924	0.256	450.45 (77,1)

Table 11 Summary of the empirical regression results in the penicillins, including the semi-empirical result obtained by Bird and Marshall<sup>210</sup> for the same structures.

Independent variables	Significance level	$\frac{n}{\sum_{i=1}^n  x_i - \hat{x}_i }$	$\frac{n}{\sum_{i=1}^n  x_i - \hat{x}_i }$ $\frac{n}{\sum_{i=1}^n  x_i - \bar{x} }$	$\frac{n}{\sum_{i=1}^n (x_i - \hat{x}_i)^2}$	$\frac{n}{\sum_{i=1}^n (x_i - \hat{x}_i)^2}$ $\frac{n}{\sum_{i=1}^n (x_i - \bar{x})^2}$	$\frac{n}{\sum_{i=1}^n  x_i - \hat{x}_i }$ $n$
Augmented pairs <sup>a</sup>	99% 10%	55.3 26.5	0.86 0.41	153.60 43.07	0.95 0.27	1.42 0.68
Augmented pairs <sup>b</sup>	99% 10%	2.68 3.13	0.04 0.05	0.41 0.53	0.00(3) 0.00(3)	0.07 0.08
$\alpha I^b$ (Agin)	-	7.08	0.11	2.12	0.01(3)	0.18

Table 12 Log (MBC) estimations for 39 local anaesthetics using the empirical regression method. Quantities  $x_i$ ,  $\hat{x}_i$ ,  $\bar{x}$  and  $n$  are defined in Table 4. A summary of the property estimations by Agin's semi-empirical method is also included.

Notes

<sup>a</sup> Property 'predictions' by the 'hold one out' technique.

<sup>b</sup> Property estimations based on the analysis of the total structure set.

Table 13 Best estimated and predicted log (MBC) values in 39 local anaesthetics by the empirical regression method.

Compound <sup>a</sup>	Observed log (MBC) <sup>b</sup>	Estimated log (MBC) <sup>c</sup>	Predicted log (MBC) <sup>d</sup>	Estimated log (MBC) (Agin) <sup>e</sup>
1	3.09	3.09	2.20*	3.08
2	2.75	2.77	2.73	2.60
3	2.60	2.47	1.71	2.37
4	2.55	2.50	2.02	2.16
5	2.40	2.26	2.19	2.15
6	2.00	2.00	0.98*	1.84
7	1.93	1.94	2.82	1.78
8	1.78	1.75	1.77	1.70
9	1.78	1.78	0.75*	1.62
10	1.77	1.64	1.73	1.67
11	1.50	1.50	2.35	1.47
12	1.40	1.31	0.94	1.53
13	1.30	1.17	1.31	1.48
14	1.30	1.33	1.45	1.04
15	1.17	1.14	1.24	1.34
16	1.20	1.24	1.27	1.26
17	1.00	1.38	1.50	1.55
18	1.00	1.10	1.25	1.16
19	0.81	0.81	0.89*	0.57
20	0.56	0.73	0.76	0.83
21	0.47	0.47	1.15*	0.66
22	0.30	0.39	0.47	0.29

Continued...



Table 13 continued

23	0.30	0.29	-0.07	0.25
24	0.20	0.23	0.19	0.38
25	0.00	-0.21	-0.39	0.19
26	0.00	0.03	0.12	-0.10
27	-0.16	-0.28	-0.35	-0.06
28	-0.52	-0.44	0.47	-0.26
29	-0.80	-0.71	-0.87	-0.79
30	-0.80	-0.80	-0.30	-0.73
31	-1.67	-1.56	-0.98	-0.62
32	-1.96	-1.93	0.00	-2.01
33	-2.80	-2.91	-3.24	-3.00
34	-2.90	-2.77	-1.64	-2.40
35	-3.20	-3.23	-1.85	-3.28
36	-3.60	-3.60	-0.47*	-3.77
37	-3.66	-3.66	-0.99*	-3.25
38	-4.00	-4.00	-0.84*	-3.93
39	-4.20	-4.33	-3.99	-4.85

Notes

<sup>a</sup> structure diagrams in Appendix 1

<sup>b</sup> taken from Agin, Hersh and Holzman<sup>209</sup>

<sup>c</sup> best estimations based on the full structure set, using augmented pairs at the 99% significance level

<sup>d</sup> best 'predictions' (based on the 'hold one out' technique), using augmented pairs at the 10% significance level.

<sup>e</sup> estimations from a regression of log (MBC) on  $\alpha I$  (Agin et al.<sup>209</sup>).

\* structures containing unique augmented pair fragments

Table 14

Independent variables	Significance level	$\sum_{i=1}^n  x_i - \hat{x}_i $	$\frac{\sum_{i=1}^n  x_i - \hat{x}_i }{\sum_{i=1}^n  x_i - \bar{x} }$	$\sum_{i=1}^n (x_i - \hat{x}_i)^2$	$\frac{\sum_{i=1}^n (x_i - \hat{x}_i)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$	$\frac{\sum_{i=1}^n  x_i - \hat{x}_i }{n}$
Simple pairs	99%	5.196	0.482	1.920	0.219	0.260
	10%	5.077	0.471	1.868	0.213	0.254
Bonded pairs	99%	4.959	0.460	2.451	0.279	0.248
	10%	5.800	0.538	2.870	0.327	0.290
Augmented atoms	99%	3.740	0.347	1.110	0.126	0.187
	10%	4.595	0.427	1.967	0.224	0.229
Octuplets	99%	5.457	0.507	2.882	0.328	0.273
	10%	5.383	0.500	3.037	0.346	0.269
Simple pairs	99%	4.260	0.396	1.443	0.164	0.213
Bonded pairs	10%	3.030	0.282	0.734	0.084	0.152
Augmented atoms	10%	2.960	0.275	0.721	0.082	0.148
Octuplets	10%	2.039	0.189	0.387	0.044	0.102
$\Sigma$ Bird & Marshall		4.045	0.376	1.119	0.127	0.202

Continued ...

Table 14 (continued) Log (B/F) estimations by the empirical regression method for a random sample of 20 of the 79 penicillins taken from Bird and Marshall. Quantities  $x_i$ ,  $\hat{x}_i$ ,  $\bar{x}$  and  $n$  are defined in Table 4, where  $x$  refers to the mean observed property value for the subset and  $n$  the number of structures in this set. A summary of Bird and Marshall's semi-empirical result for the same subset is also included.

Notes

<sup>a</sup> see Table 12

<sup>b</sup> best property estimations based on the analysis of the total set of 79 structures

Table 15 Best estimated and predicted log (B/F) values in the random sample of 20 penicillin structures, by the empirical regression method.

Compound <sup>a</sup>	Observed log (B/F) <sup>b</sup>	Estimated log (B/F) <sup>c</sup>	Predicted log (B/F) <sup>d</sup>	Estimated log (B/F) (B&M) <sup>e</sup>
1	-0.659	-0.774	-0.907*	-0.628
4	1.144	1.316	1.363	1.656
8	-0.052	-0.135	-0.199	-0.218
9	-0.602	-0.329	-0.386	-0.374
18	0.525	0.346	0.232	0.363
23	1.195	1.212	1.084	0.987
24	1.195	1.212	1.084	0.963
27	0.176	-0.352	-0.907	0.026
29	0.664	0.497	0.522	0.494
32	-0.695	-0.601	-0.514	-0.457
46	0.176	0.066	0.148*	0.260
48	0.644	0.783	0.717	0.646
55	1.380	1.111	1.064	0.934
56	1.261	1.111	1.074	0.987
67	0.602	0.693	0.562	0.855
68	0.921	0.887	1.177	1.099
70	1.574	1.274	1.238	1.251
76	0.122	0.203	0.033	0.202
78	0.362	0.397	0.453	0.446
79	0.466	0.591	0.536	0.690

Continued ...

Notes to Table 15

- a structure diagrams in Appendix 1.
- b taken from Bird and Marshall.<sup>210</sup>
- c best estimations based on the full structure set, using augmented atoms at the 10% significance level.
- d best 'predictions' based on the 'hold one out' technique, using augmented atoms at the 99% level.
- e estimations from a regression of  $\log (B/F)$  on  $\Sigma \Pi$  (Bird and Marshall<sup>210</sup>).
- \* structures containing unique augmented atom fragments.

Table 16 Results of the empirical regression analysis on 39 local anaesthetics using augmented pairs at the 10% significance level.

Structural feature	Regression coefficient	Student t (20 degrees of freedom )
$\emptyset$ C - C1	-0.80	11.12
1C - C1 (chain)	-0.51	18.03
1C - C2 (chain)	-1.08	13.33
2C - C2 (chain)	-1.10	18.36
1C - C3 (chain)	-1.95	17.90
1C * C1	-0.14	5.70
1C * C2	-0.31	9.43
2C * C2	-0.38	7.32
$\emptyset$ C - N2	-0.53	6.38
2C - N2 (chain)	-0.68	3.83
2N = $\emptyset$		
2C - N2 (ring)	-0.58	5.60
$\emptyset$ C - C $\emptyset$	0.74	3.96
1C - $\emptyset$	1.24	10.13
1C - O1 (chain)	0.61	5.60
2C - $\emptyset$	0.21	2.46
2C - O1 (chain)	0.25	2.04
2C - O1 (ring)		
2N - O1 (ring)	-0.57	3.04
1N - N2 (ring)		
2C - C1	-0.28	4.53
regression constant	2.35	25.88

$\emptyset$  = zero

Continued ...

Note to Table 16

The fragments are represented in the form  $n_a A - B_{n_b}$  where  $n_a$  and  $n_b$  are the numbers of non-hydrogen atoms bonded to atoms A and B respectively. The terms ring and chain given after fragments indicate the positions of atoms A and B in the structure, in cases where these are not clear

Table 17 Results of the empirical regression analysis on 79 penicillins using augmented atoms at the 10% significance level

Structural feature	Regression coefficient	Student t (65 degrees of freedom)
$\begin{array}{c} \text{C} - \text{C} - \text{C} \\   \\ \text{Br} \end{array}$	-0.492	1.83
$\begin{array}{c} \text{H}_2\text{N} - \text{S} \\   \\ \text{C} \\ * \\ \text{C} * \text{C} - \text{S} \\ * \\ \text{C} \end{array}$	-0.463	1.73
$\begin{array}{c} \text{N} \\   \\ \text{O} = \text{S} = \text{O} \\   \\ \text{C} \end{array}$		
$\text{O} = \text{S}$		
$\text{Br} - \text{C}$	0.854	5.22
$\text{C} * \text{S} * \text{C}$		
$\begin{array}{c} \text{S} * \text{C} - \text{C} \\ * \\ \text{C} \end{array}$	0.315	2.75
$\text{F} - \text{C}$		
$\begin{array}{c} \text{C} * \text{C} - \text{F} \\ * \\ \text{C} \end{array}$	0.255	3.16
$\text{H}_2\text{N} - \text{C}$	-0.470	7.14
$\text{Cl} - \text{C}$	0.573	17.45
$\text{C} * \text{C} * \text{C}$	0.137	2.46
$\begin{array}{c} \text{O} \\   \\ \text{C} - \text{C} - \text{C} \end{array}$	0.251	9.26

Continued ...



Table 17 continued

$\begin{array}{c} \text{C} * \text{C} * \text{C} \\   \\ \text{C} \end{array}$	-0.141	2.41
C - N - C	-0.828	5.36
O = C		
HO - C	-0.334	2.84
H <sub>3</sub> C - C	0.194	4.73
C * C * C	0.245	14.49
regression constant	excluded from the regression	

Note to Tables 16 and 17

Delocalized ring bonds are denoted by asterisks. All other bonds indicated in the penicillins are present in chains. The perfectly correlated fragments in each sample (intra group correlation coefficients of + 1) are bracketed, and only one fragment from each group is included in the calculations.

Method	Structural feature	$\sum_{i=1}^n  x_i - \hat{x}_i $	$\frac{\sum_{i=1}^n  x_i - \hat{x}_i }{\sum_{i=1}^n  x_i - \bar{x} }$	$\sum_{i=1}^n (x_i - \hat{x}_i)^2$	$\frac{\sum_{i=1}^n (x_i - \hat{x}_i)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$	$\frac{\sum_{i=1}^n  x_i - \hat{x}_i }{n}$
Regression	Augmented pairs <sup>a</sup>	26.50	0.41	43.07	0.27	0.68
Nearest neighbour(s)	Augmented pairs <sup>b</sup>	33.71	0.52	76.79	0.48	0.86
Classification	Augmented pairs <sup>c</sup>	41.08	0.64	73.56	0.46	1.05
Nearest neighbour(s)	Atoms <sup>c</sup>	24.70	0.38	21.51	0.13	0.63
Classification	Atoms <sup>c</sup>	28.87	0.45	34.45	0.21	0.71

Table 18 A summary of the best prediction results in the anaesthetics by the nearest neighbour, classification and empirical regression methods

Notes

- <sup>a</sup> at the 10% significance level  
<sup>b</sup> using the simple distance coefficient  
<sup>c</sup> using Dice's coefficient

Table 19

Method	Structural feature	$\sum_{i=1}^n  x_i - \hat{x}_i $	$\frac{\sum_{i=1}^n  x_i - \hat{x}_i }{\sum_{i=1}^n  x_i - \bar{x} }$	$\sum_{i=1}^n (x_i - \hat{x}_i)^2$	$\frac{\sum_{i=1}^n (x_i - \hat{x}_i)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$	$\frac{\sum_{i=1}^n  x_i - \hat{x}_i }{n}$
Regression	Simple pairs <sup>a</sup>	5.077	0.471	1.868	0.213	0.254
	bonded pairs <sup>b</sup>	4.959	0.460	2.451	0.279	0.248
	octuplets <sup>b</sup>	5.457	0.507	2.882	0.328	0.273
	augmented <sup>b</sup> atoms	3.740	0.347	1.110	0.126	0.187
Nearest neighbour(s)	Simple pairs <sup>c</sup>	5.227	0.485	2.056	0.234	0.261
	bonded pairs <sup>d</sup>	6.655	0.618	3.212	0.365	0.333
	octuplets <sup>d</sup>	6.149	0.571	3.393	0.387	0.307
	augmented <sup>c</sup> atoms	5.060	0.470	2.314	0.264	0.253
Classification	Simple pairs <sup>c</sup>	6.731	0.625	4.094	0.466	0.337
	bonded pairs <sup>d</sup>	7.657	0.711	3.875	0.441	0.383

Continued ...

(Classification)	octuplets <sup>d</sup>	6.049	0.562	3.029	0.345	0.302
	augmented <sup>c</sup> atoms	4.745	0.441	1.589	0.181	0.237

Table 19 (continued) A summary of the best prediction results in the sample of 20 penicillin structures, by the nearest neighbour, classification and empirical regression methods.

Notes

<sup>a</sup> at the 10% significance level

<sup>b</sup> at the 99% significance level

<sup>c</sup> using the simple distance coefficient

<sup>d</sup> using Dice's coefficient

Independent variables		Significance level	Variables included in regression	Degrees of freedom	Multiple correlation coefficient	Residual error	F statistic
Type	No						
Simple pairs	27	99%	26 + const	103	0.848	0.489	10.136 (103,26)
		10%	13 + const	116	0.817	0.502	17.873 (116,13)

Table 20 Summary of the empirical regression results for a group of 130 mixed structures from the literature with known partition coefficients. (see Appendix III).

Independent variables	Variables included in regression	Degrees of freedom	Multiple correlation coefficient	Residual error	Regression coefficient	Regression constant	F statistic
log P estimated <sup>b</sup>	1 + C	77	0.786	0.414	0.6619	1.1726	124.465
log P estimated <sup>c</sup>	1 + C	77	0.766	0.431	0.6502	-0.5264	109.441

Table 21 Summary of the regressions of observed log (B/F) values on estimated partition coefficients<sup>a</sup> in 79 penicillins

Notes

<sup>a</sup> estimated from the fragment contributions obtained from the analysis of 130 structures taken from the literature with known partition coefficients (see Table 20).

<sup>b</sup> fragment II values obtained at the 99% significance level.

<sup>c</sup> fragment II values obtained at the 10% significance level.

Table 22 Serum binding predictions for a group of 18 penicillins\*, obtained by the semi-empirical regression method described in Appendix III.

Compound	Observed B	Predicted B	$\Delta B$
1	54	42	12
2	47	55	8
3	60	74	14
4*	64	60	4
5*	22	95	13
6	89	92	3
7	67	83	16
8*	11	54	43
9	5	7	2
10	49	99	50
11	45	51	6
12	88	70	12
13	94	74	20
14	50	31	19
15	26	10	16
16	50	15	35
17	86	50	36
18	95	9	86

\* Structures supplied by A.E. Bird, Beecham Pharmaceuticals, Chemotherapeutic Research Centre, Brockam Park, Betchworth, Surrey.



## Notes to Figures

1.

### Fragment Key (to Figures 19 to 24)

<u>Code</u>	<u>Fragment Type</u>
A	Atoms
SP	Simple pairs
AP	Augmented pairs
BP	Bonded pairs
OC	Octuplets
AA	Augmented atoms

2.

### Bond Key (to Figure 44)

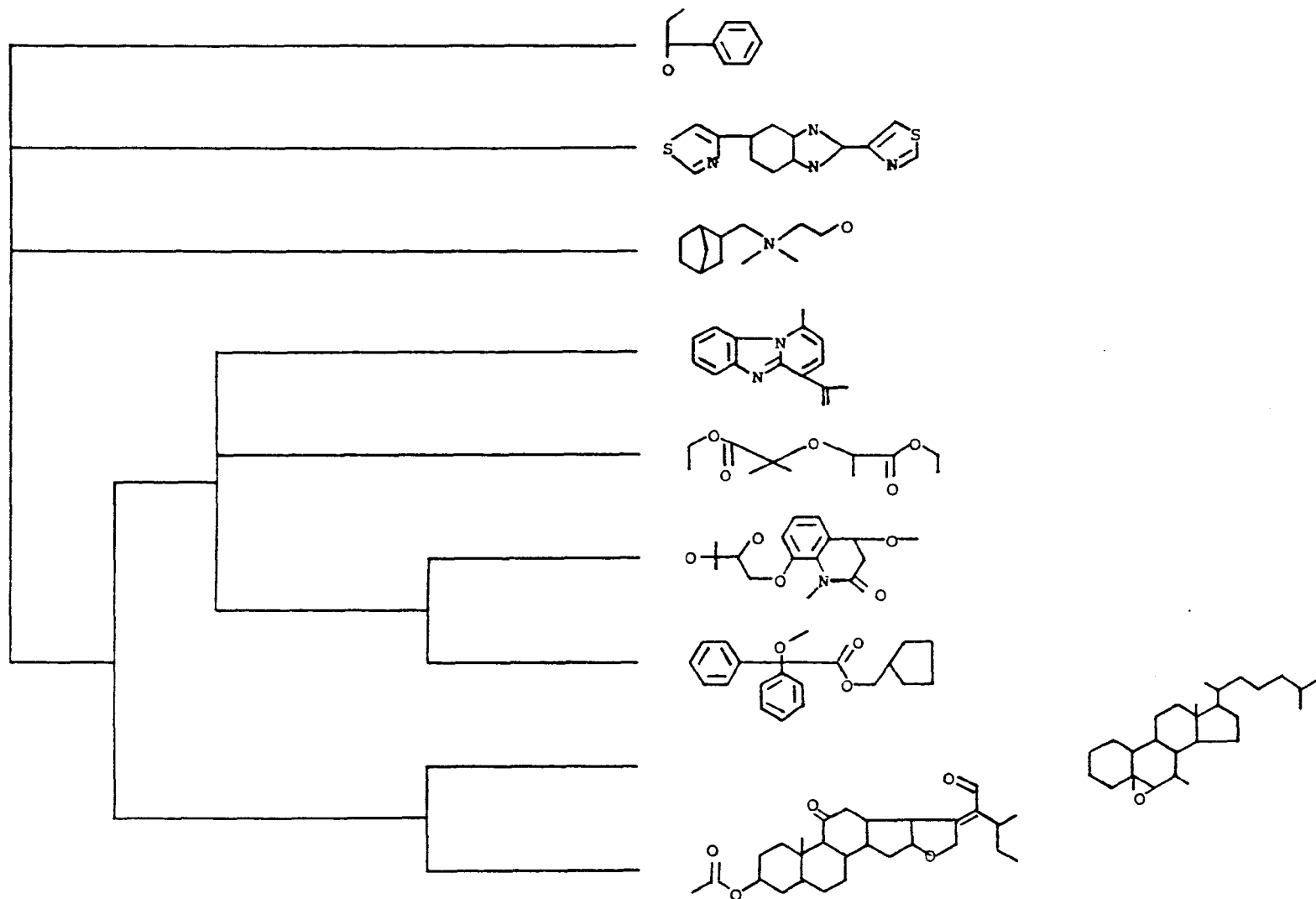
<u>Code</u>	<u>Bond Type</u>
1	Acyclic single bond
2	Acyclic double bond
3	not used
4	Acyclic triple bond
5	Cyclic single bond
6	Cyclic double bond
7	Cyclic aromatic bond
8	Cyclic triple bond

3.

All classifications were carried out using the single-link clustering method<sup>46</sup>.

4.

Delocalised ring bonds in fragment definitions are represented by asterisks. (Other bonds indicated are self explanatory).



**Figure 1** Dendrogram showing the classification obtained for a CAS registry file sample using structure representation (i) and a simple matching coefficient based on the number of attributes shared between structures

Figure 2 Augmented atom fragments occurring in the amino acids asparagine and glutamine.

Structure	Fragments
$  \begin{array}{ccccccc}  \text{H}_2\text{N} & \text{C} & \text{CH}_2 & \text{CH} & \text{COOH} \\  &    & &   & \\  & \text{O} & & \text{NH}_2 &   \end{array}  $	$  \begin{array}{l}  \text{O} - \text{C} \\  \text{N} - \text{C} \quad \times 2 \\  \text{N} - \text{C} - \text{C} \\  \quad    \\  \quad \text{O} \\  \text{C} = \text{O} \quad \times 2 \\  \text{C} - \text{C} - \text{C} \\  \text{C} - \text{C} - \text{C} \\  \quad   \\  \quad \text{N} \\  \text{C} - \text{C} - \text{C} \\  \quad    \\  \quad \text{O}  \end{array}  $
$  \begin{array}{ccccccc}  \text{H}_2\text{N} & \text{C} & \text{CH}_2 & \text{CH}_2 & \text{CH} & \text{COOH} \\  &    & & &   & \\  & \text{O} & & & \text{NH}_2 &   \end{array}  $	$  \begin{array}{l}  \text{O} - \text{C} \\  \text{N} - \text{C} \quad \times 2 \\  \text{N} - \text{C} - \text{C} \\  \quad    \\  \quad \text{O} \\  \text{C} = \text{O} \quad \times 2 \\  \text{C} - \text{C} - \text{C} \times 2 \\  \text{C} - \text{C} - \text{C} \\  \quad   \\  \quad \text{N} \\  \text{C} - \text{C} - \text{C} \\  \quad    \\  \quad \text{O}  \end{array}  $

Using structure representation (ii),  $a = 9$  (attributes common to both structures),  $b = 0$  (attributes only in asparagine),  $c = 1$  (attributes only in glutamine) and  $d = 35$  (attributes absent from both structures for the given structure set). Therefore  $\phi = 0.935$  (see next figure and Table 1).

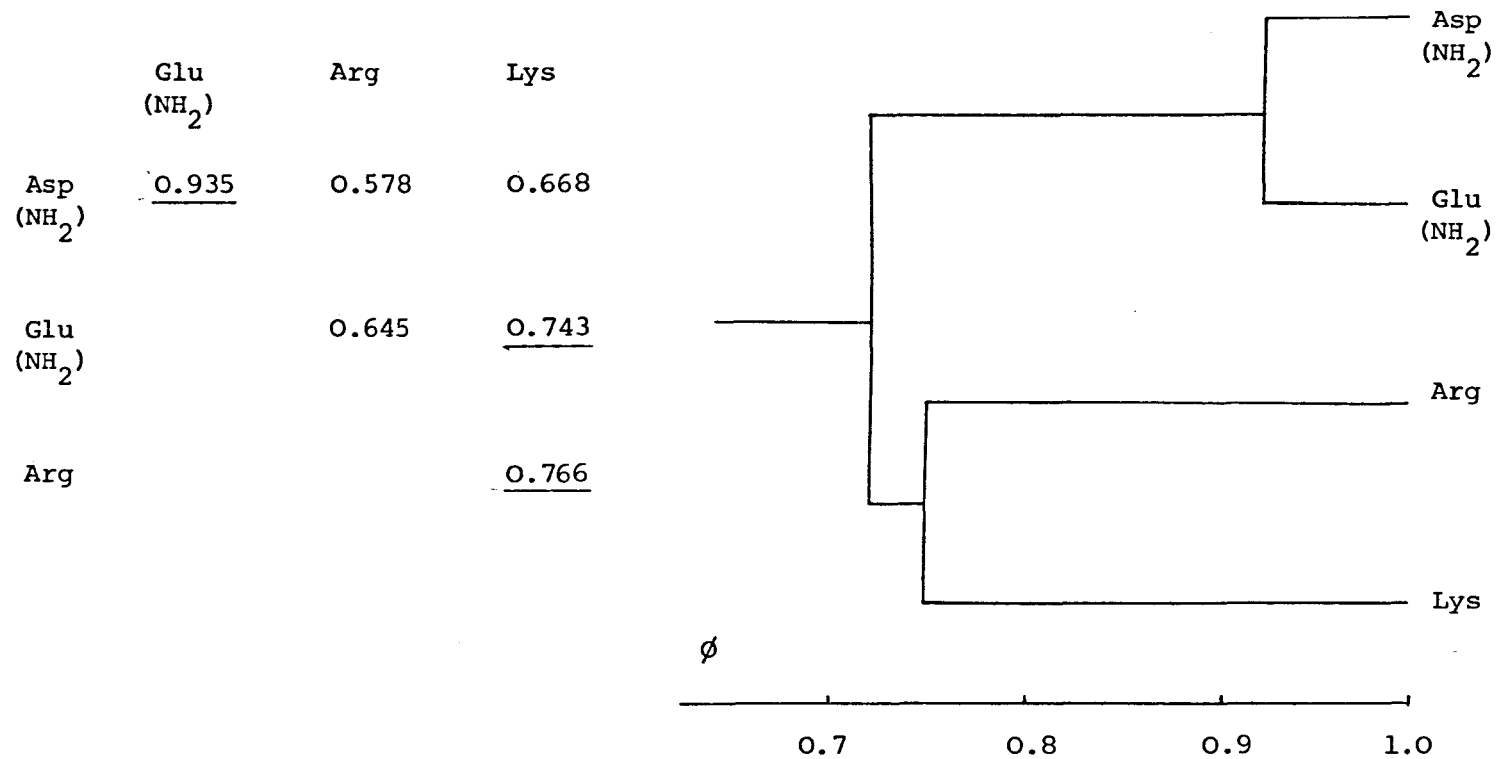


Figure 3 Initial single-link clusters formed between the amino acids, asparagine, glutamine, arginine and lysine, using  $\phi$  and structure representation (ii).

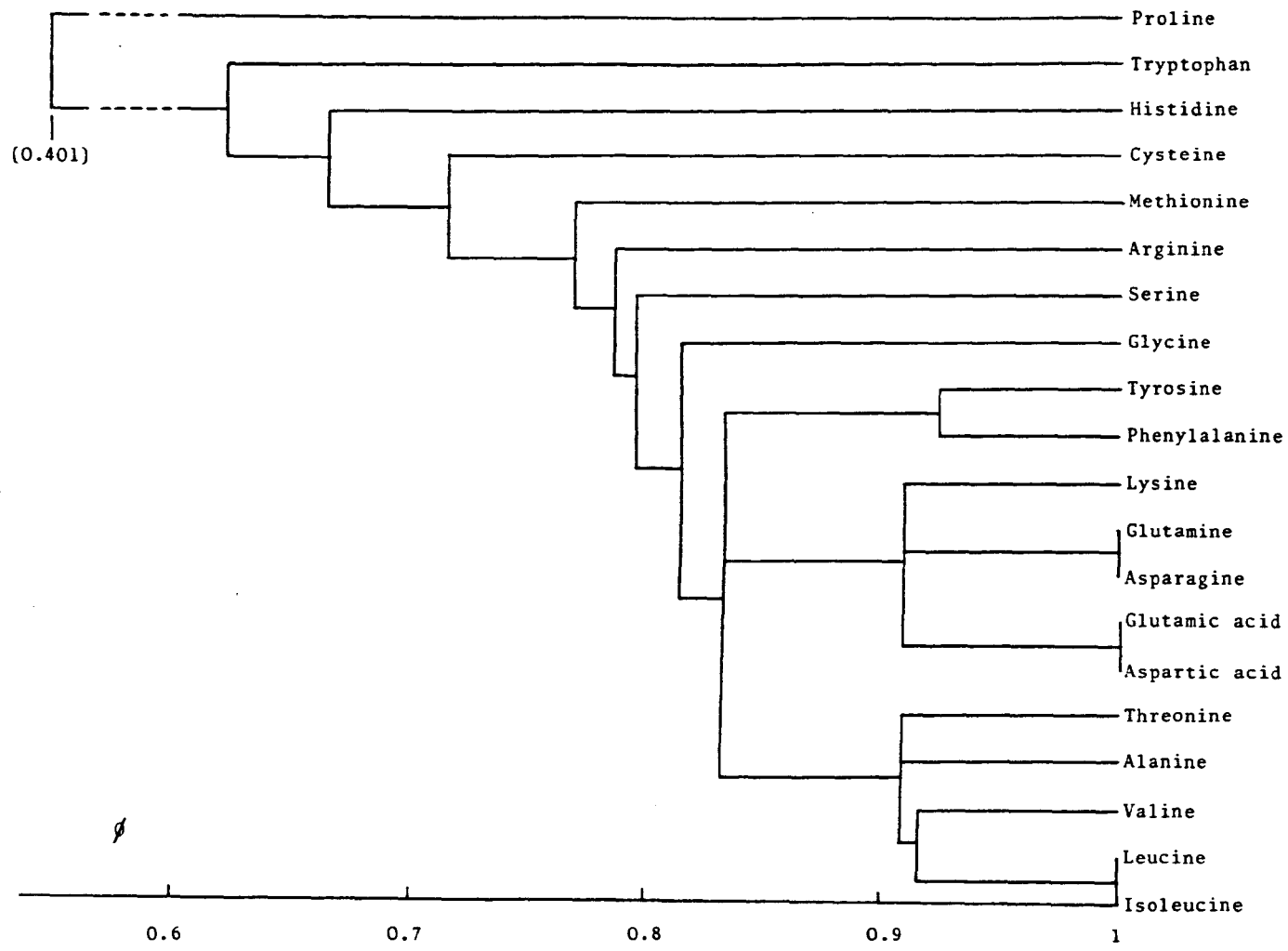


Figure 4 Dendrogram showing the classification obtained for 20 amino acids using  $\phi$  and structure representation (i)

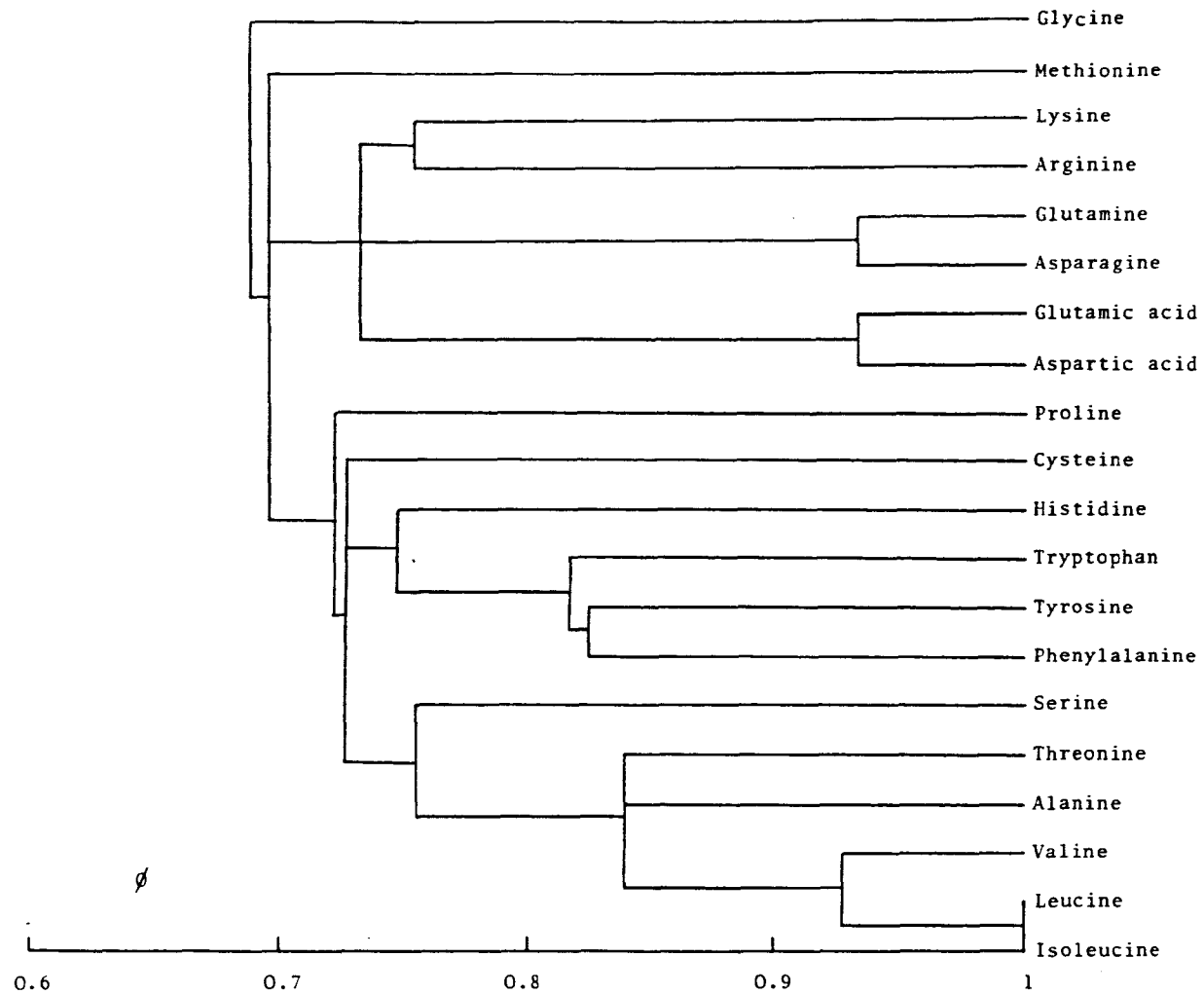
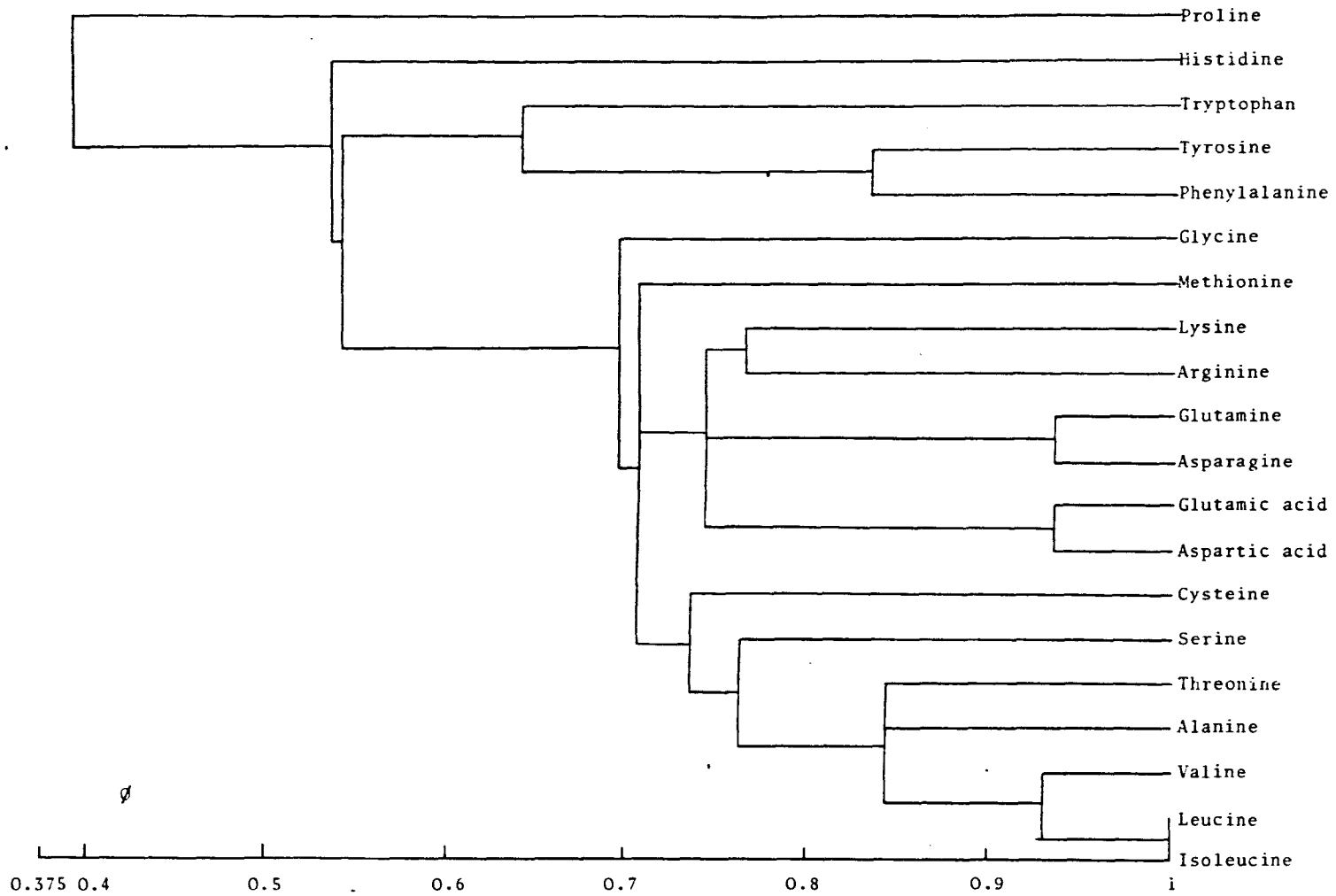
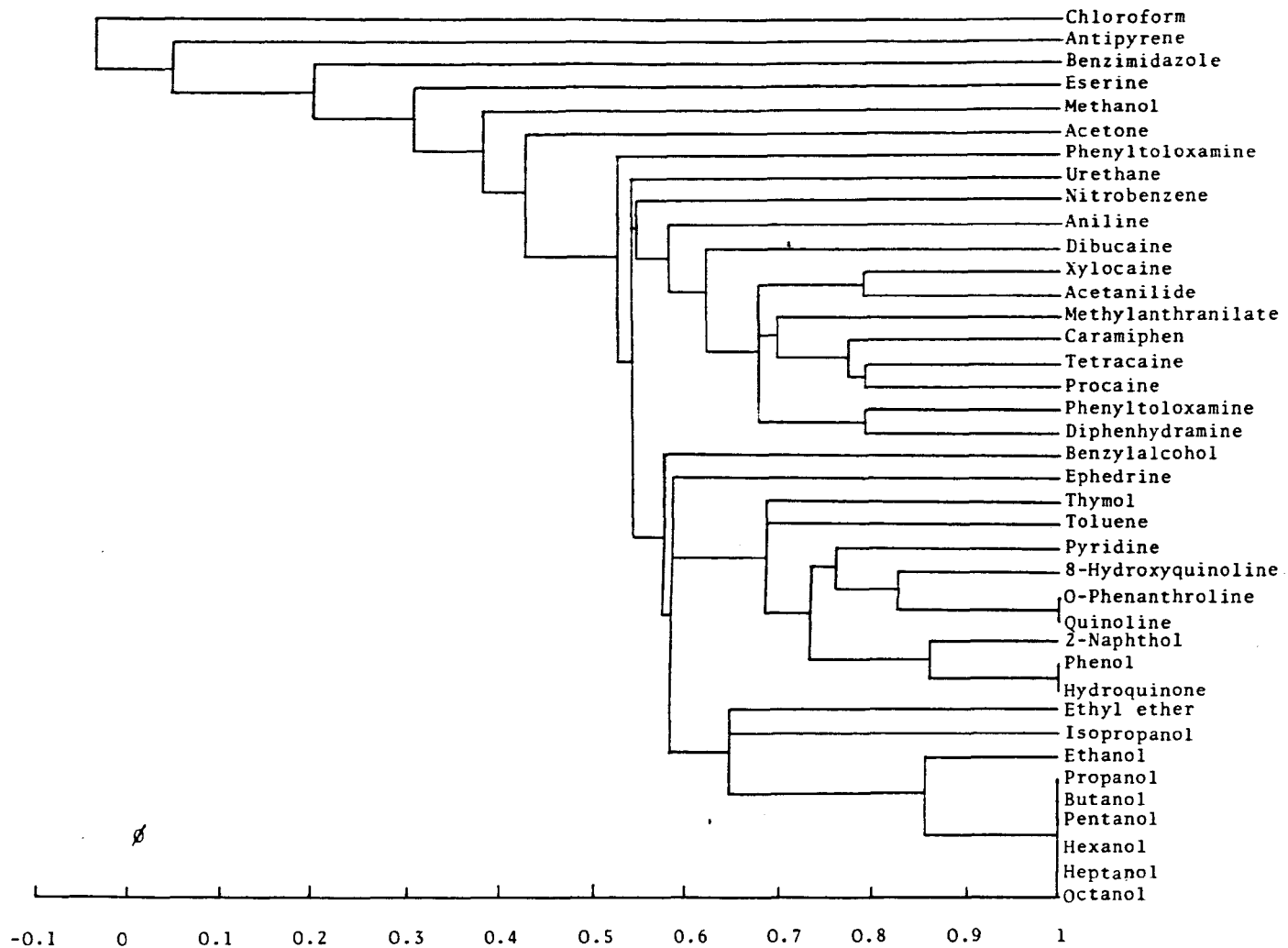


Figure 5 Dendrogram showing the classification obtained for 20 amino acids using  $\phi$  and structure representation (iii)

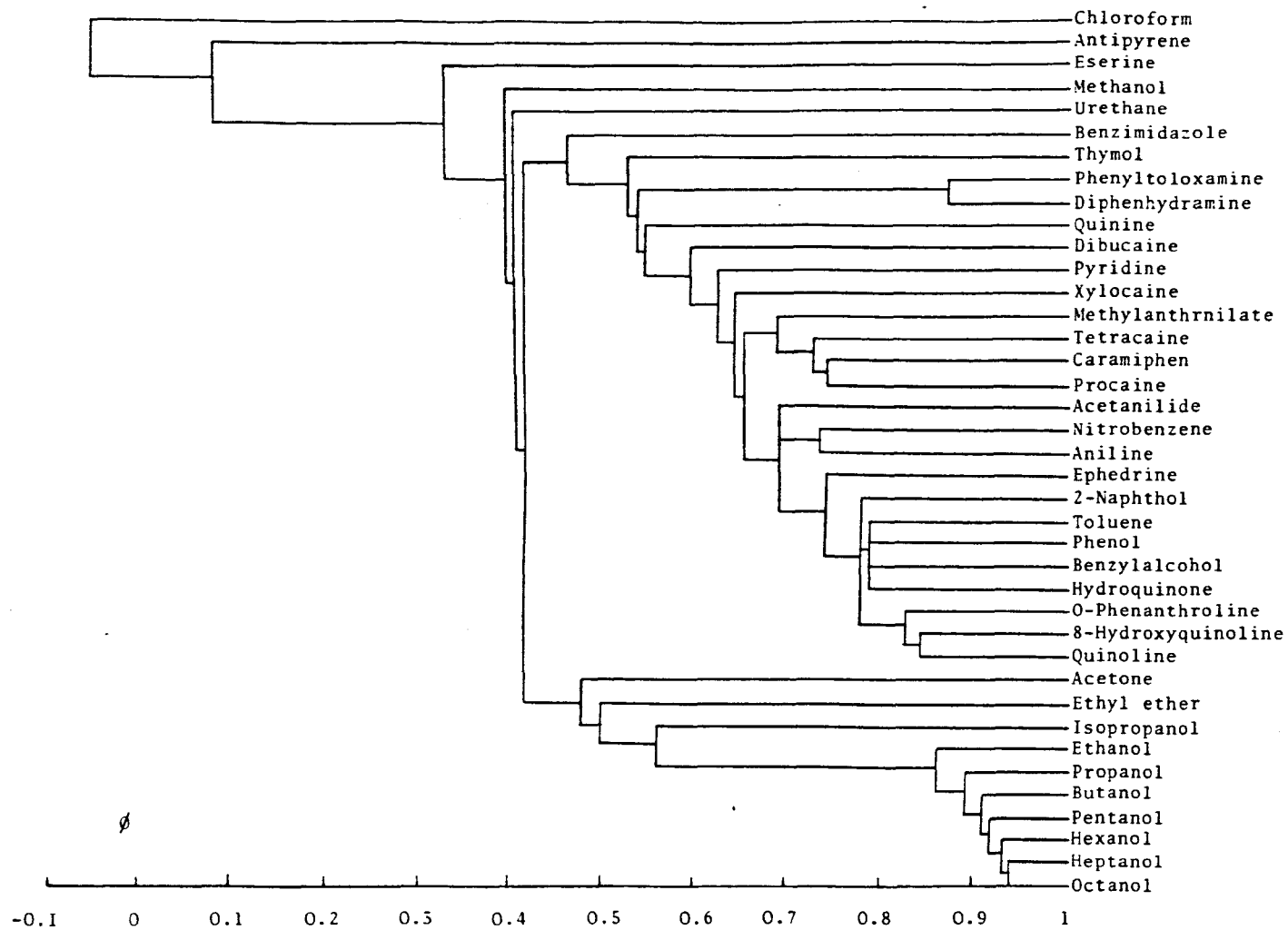


**Figure 6** Dendrogram showing the classification obtained for 20 amino acids using  $\emptyset$  and structure representation (ii)



**Figure 7** Dendrogram showing the classification obtained for 39 local anaesthetics using  $\phi$  and structure representation (i)





**Figure 8** Dendrogram showing the classification obtained for 39 local anaesthetics using  $\phi$  and structure representation (ii)

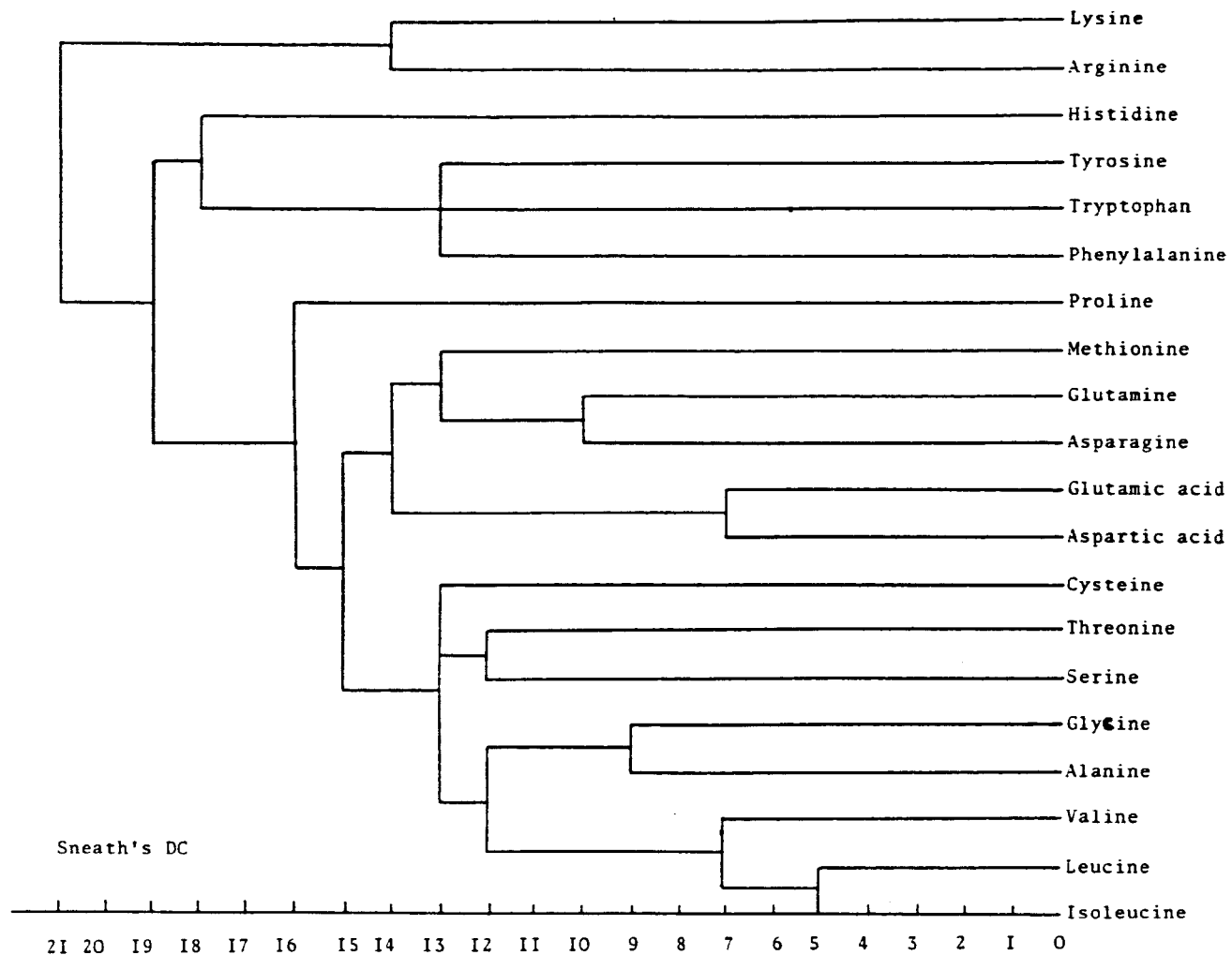


Figure 9 Dendrogram showing the classification obtained for 20 amino acids using the DC values derived by Sneath on the basis of structural and physicochemical descriptors combined <sup>10</sup>.

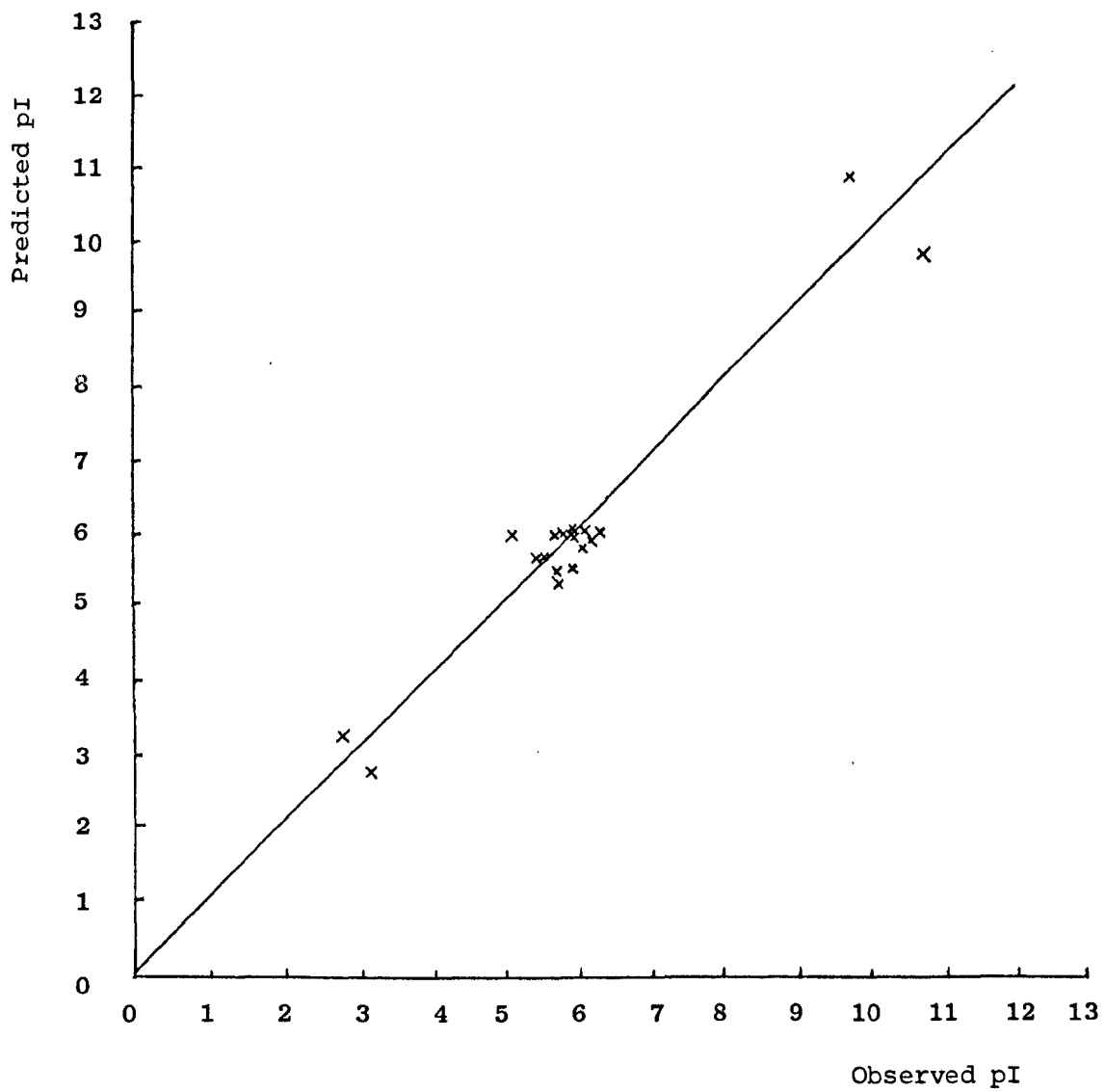


Figure 9A Observed against 'predicted' pI values in 20 amino acids using the classification based on  $\phi$  and structure representation (ii).

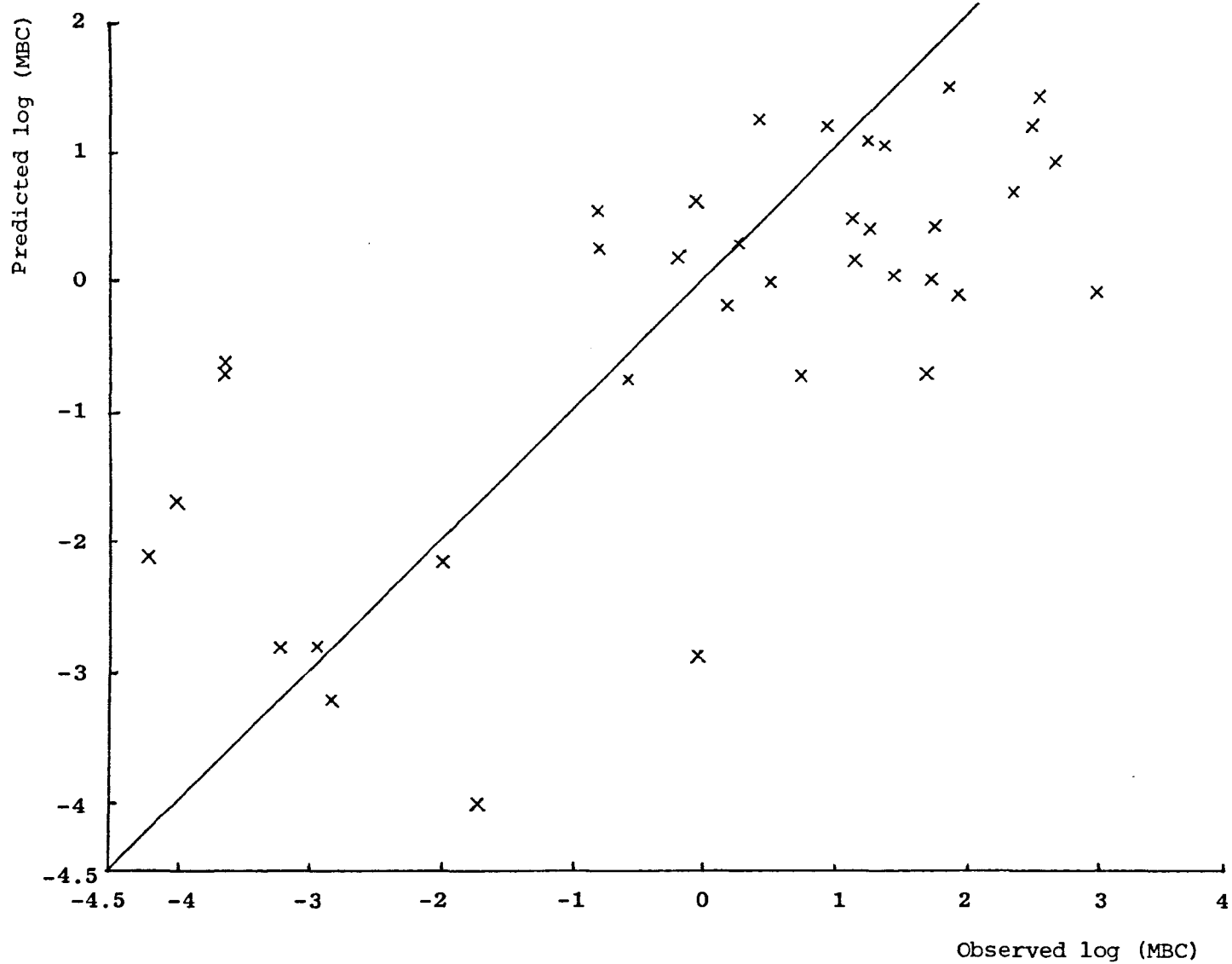


Figure 9B Observed against 'predicted' log (MBC) values in 39 local anaesthetics using the classification based on Dice's SC and structure representation (ii).

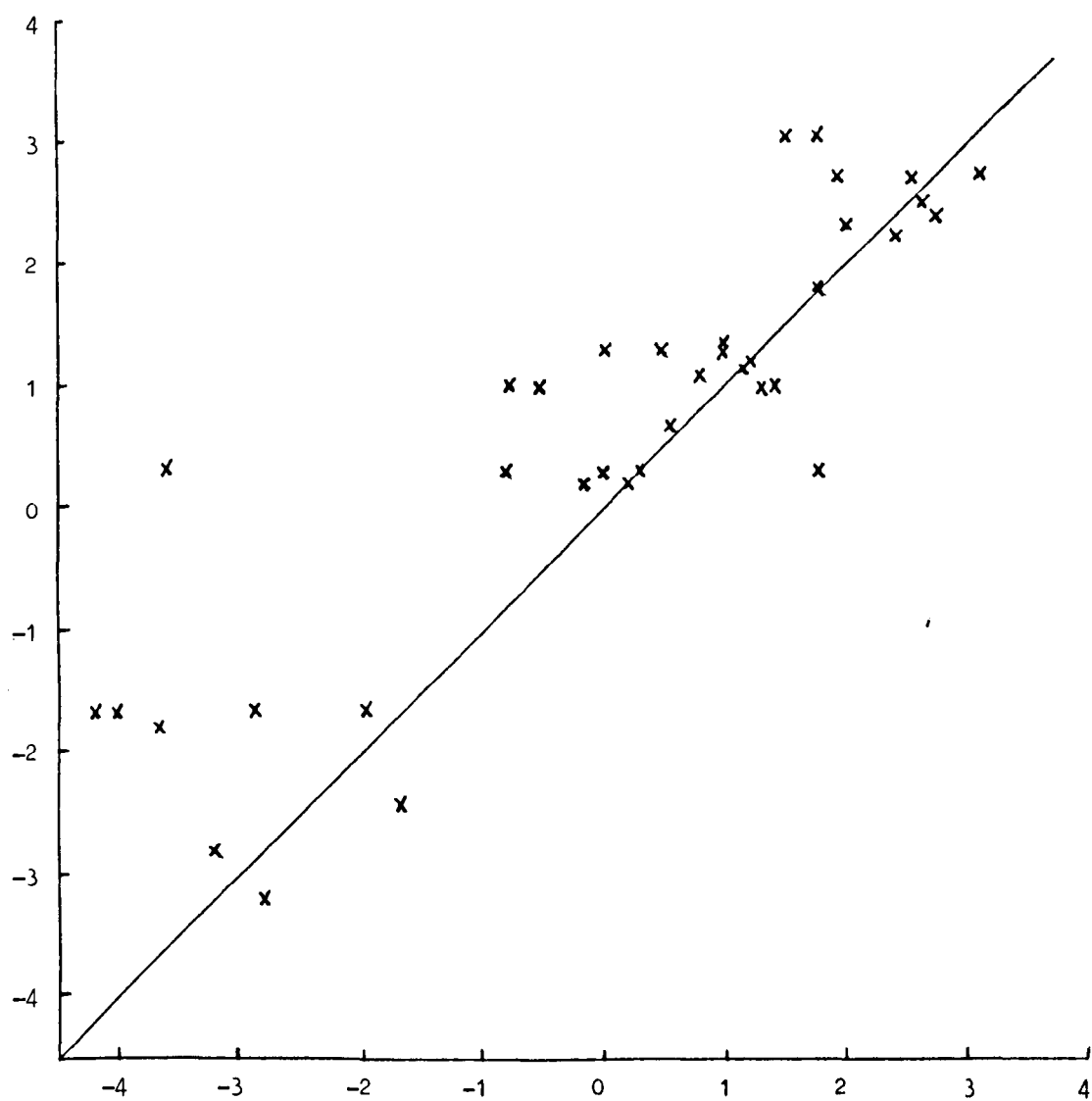


Figure 10 Observed against 'predicted' log (MBC) values in 39 local anaesthetics using highest associations based on the simple distance coefficient and augmented atom descriptors

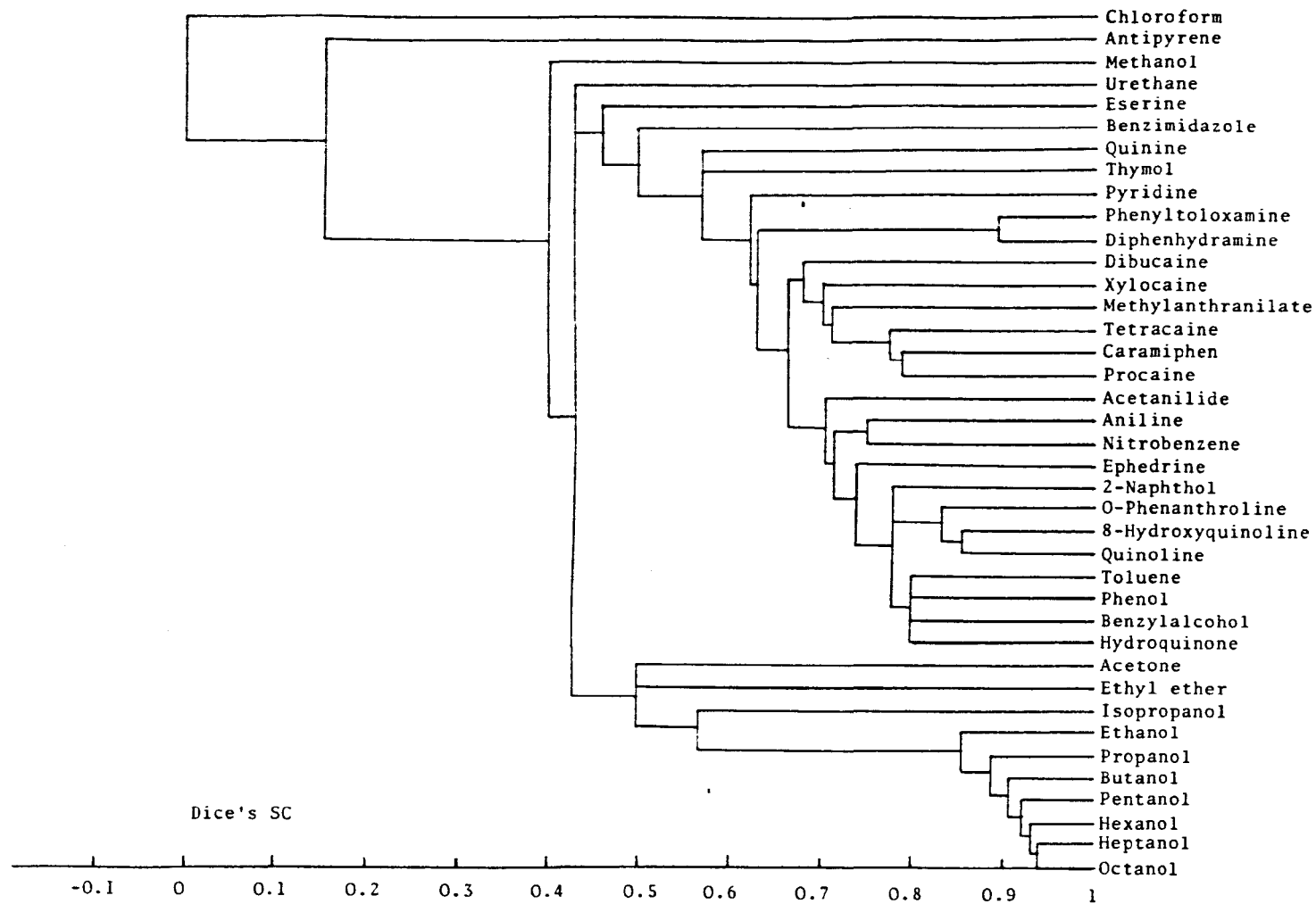
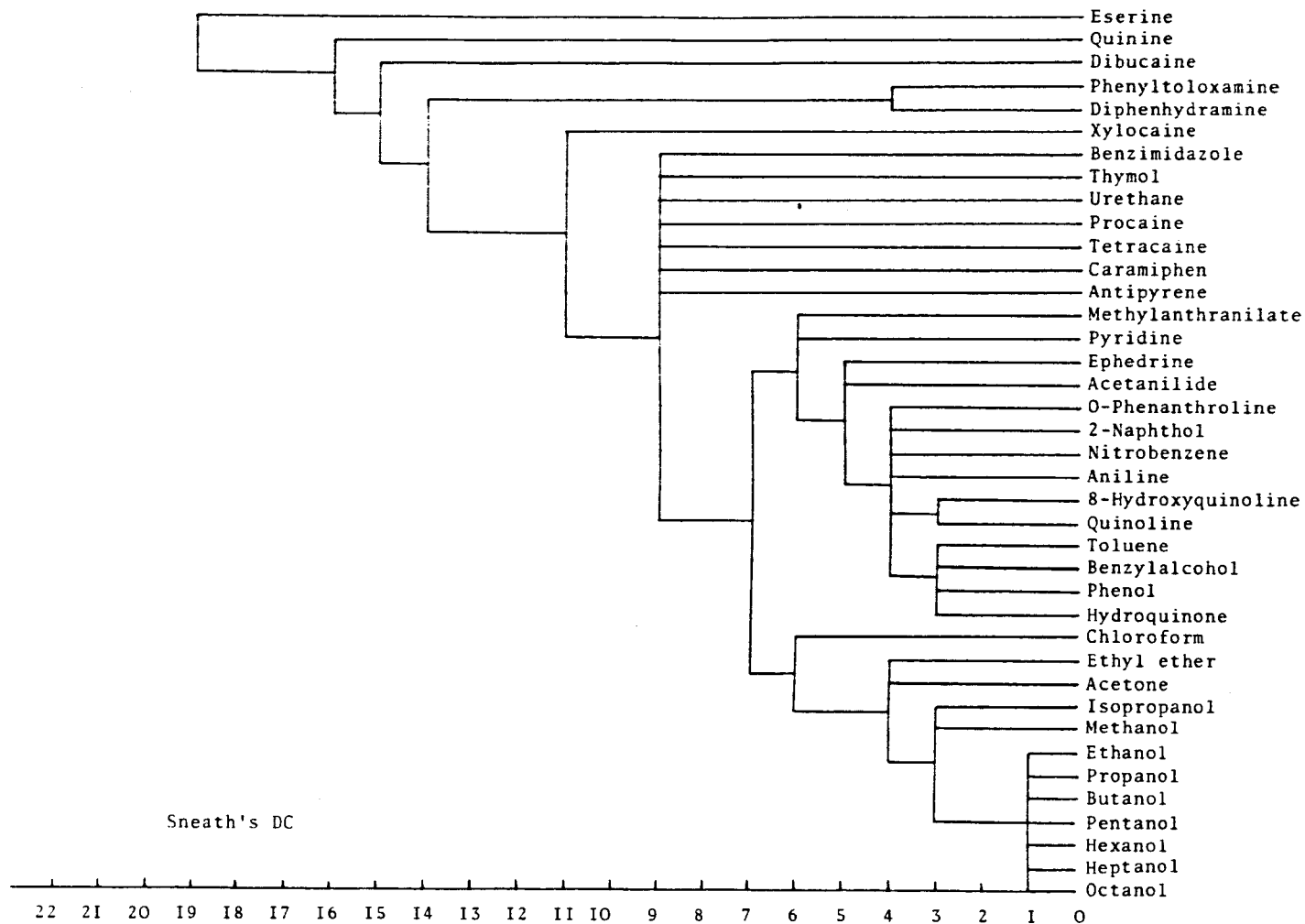
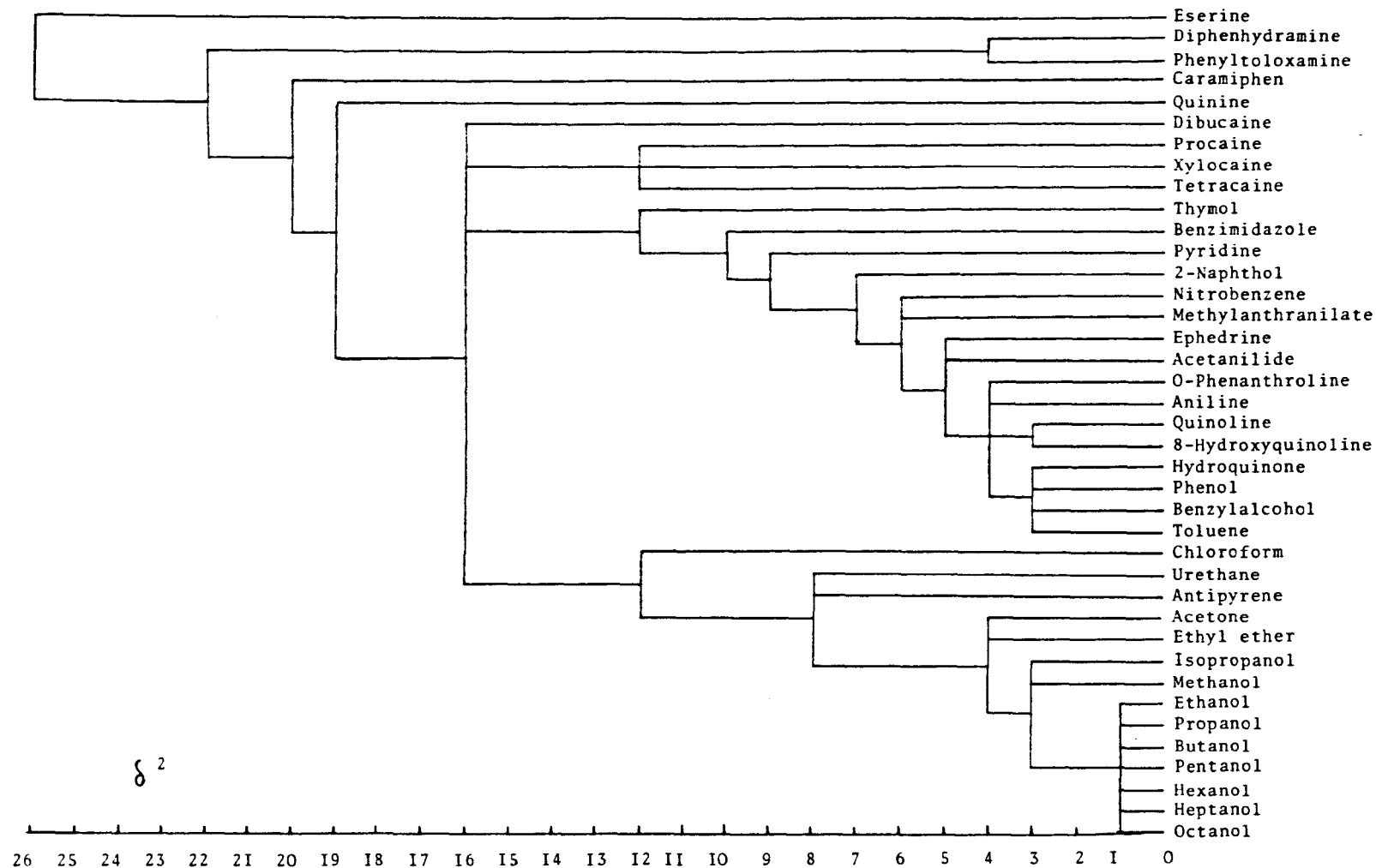


Figure 11 Dendrogram showing the classification obtained for 39 local anaesthetics using Dice's coefficient and structure representation (ii)



**Figure 12** Dendrogram showing the classification obtained for 39 local anaesthetics using Sneath's DC and structure representation (ii)



**Figure 13** Dendrogram showing the classification obtained for 39 local anaesthetics using the simple distance coefficient and structure representation (ii)'



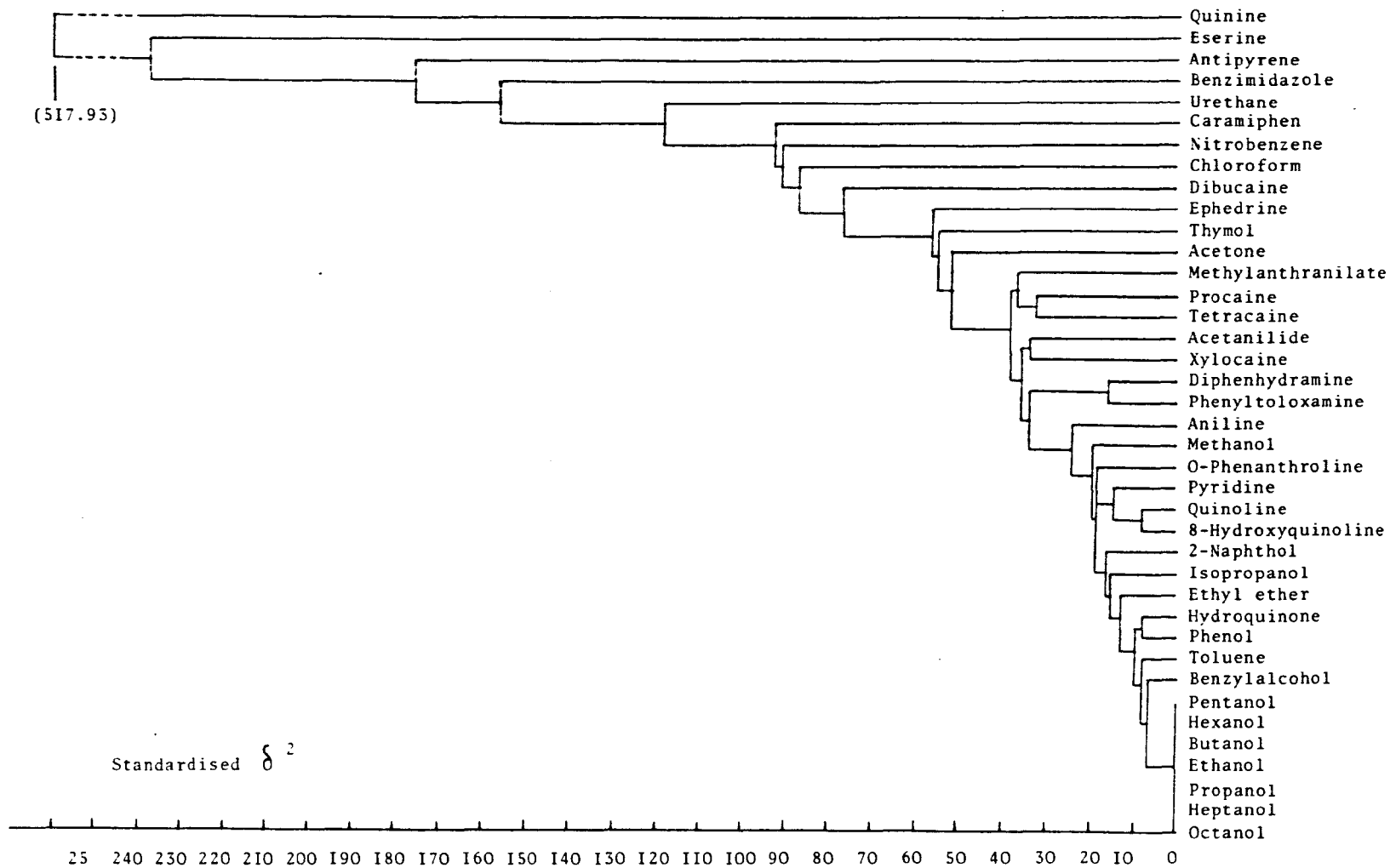
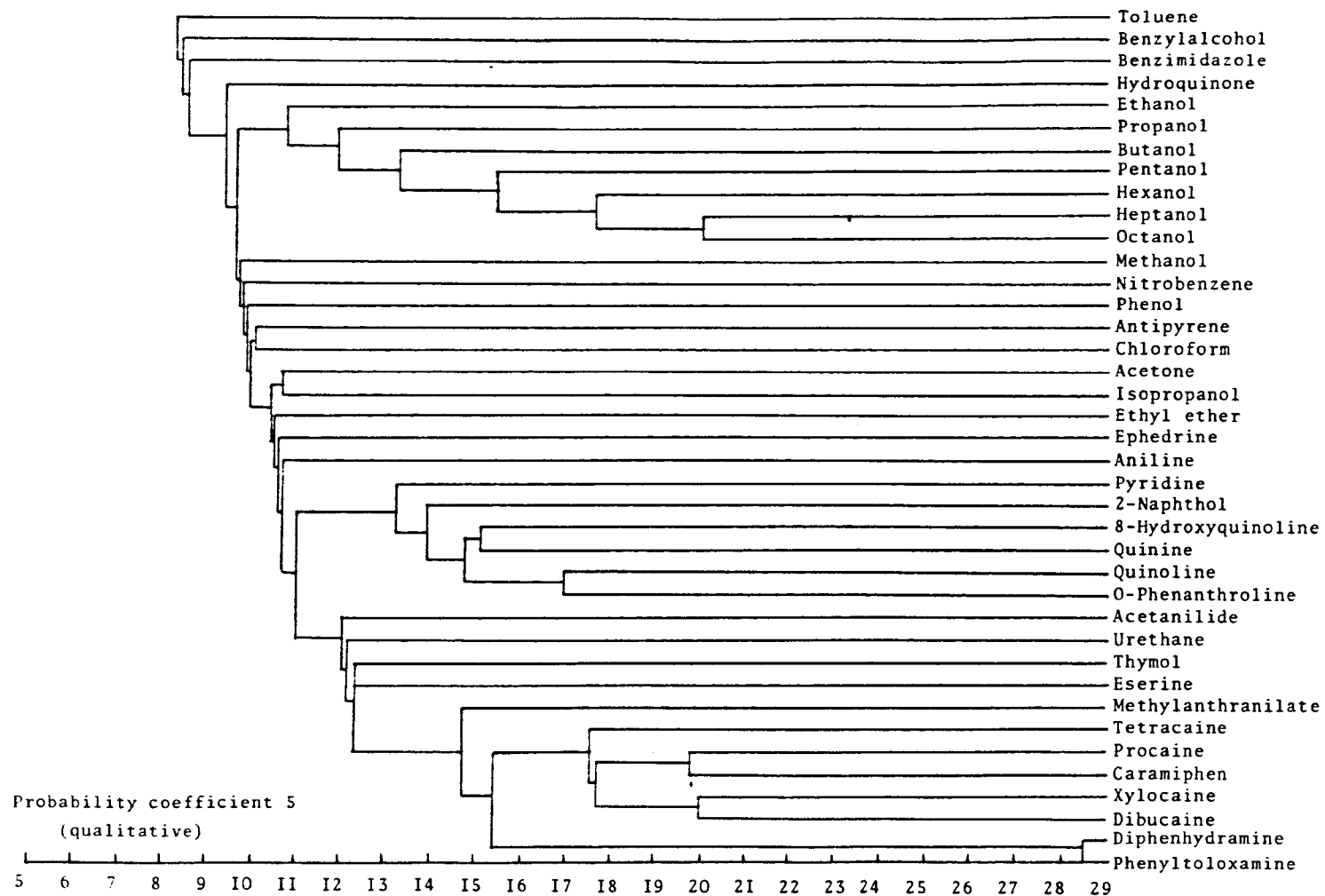


Figure 14 Dendrogram showing the classification obtained for 39 local anaesthetics using the standardised distance coefficient and structure representation (ii)'



**Figure 15** Dendrogram showing the classification obtained for 39 local anaesthetics using a probability coefficient based on structure representation (ii)

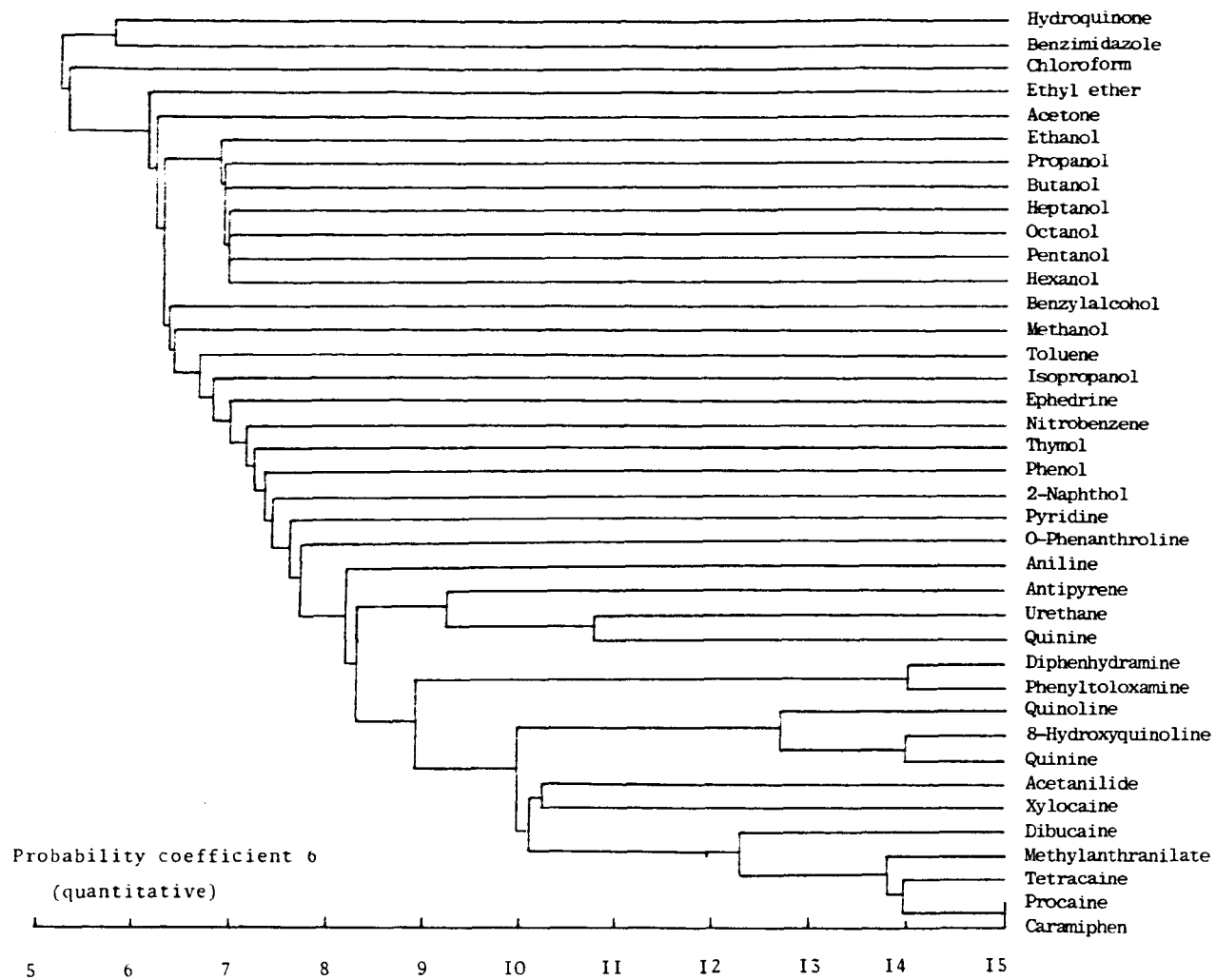
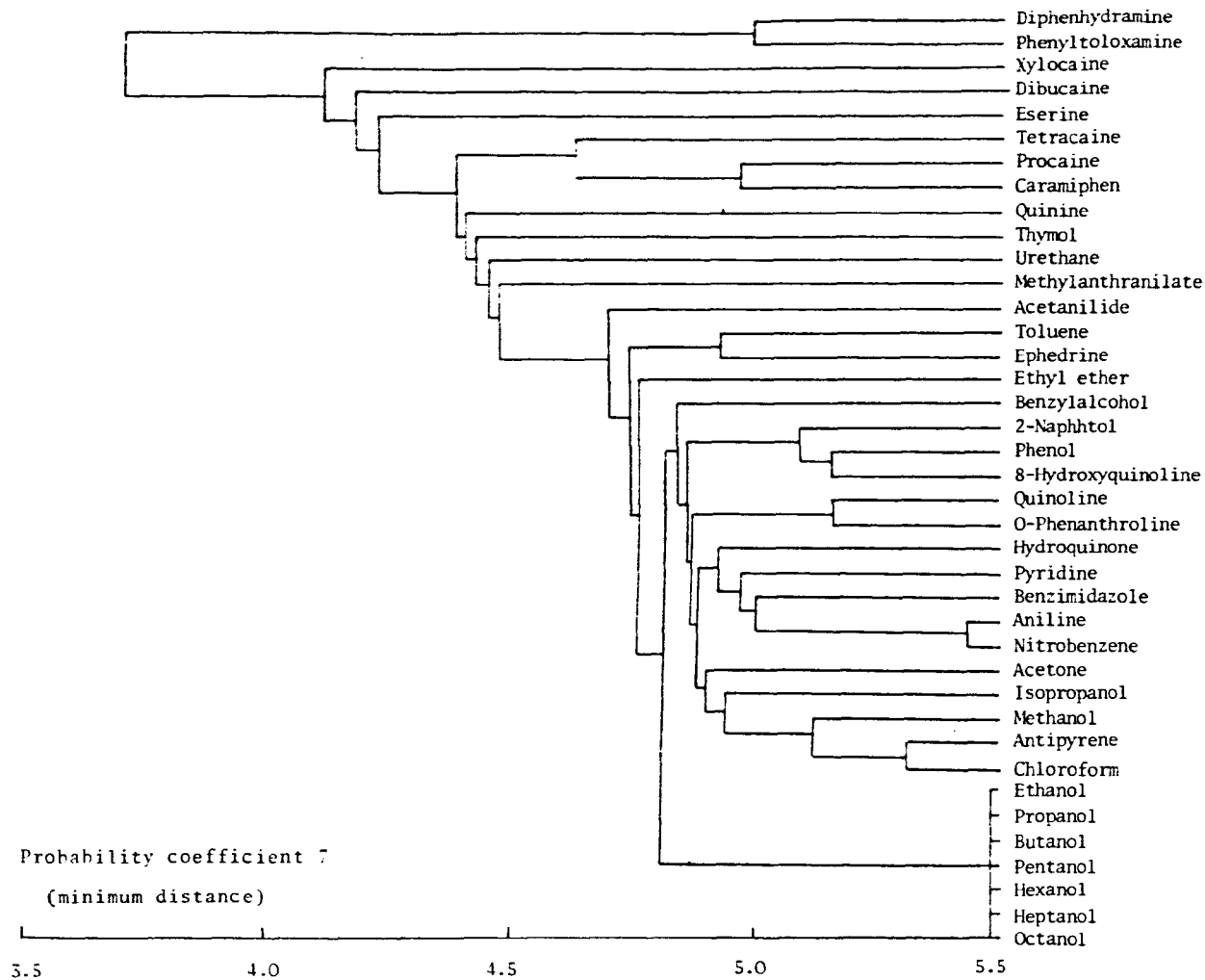


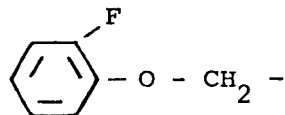
Figure 16 Dendrogram showing the classification obtained for 39 local anaesthetics using a probability coefficient based on structure representation (ii)'



**Figure 17** Dendrogram showing the classification obtained for 39 local anaesthetics using a 'minimum distance' probability coefficient (see text) based on structure representation (ii)

Figure 18 Augmented atom, simple pair, augmented pair, bonded pair and octuplet fragments occurring in three penicillin side chains, showing the distinction possible between ortho ring derivatives and meta and para derivatives when using larger bond - centred fragments.

Side chain 1

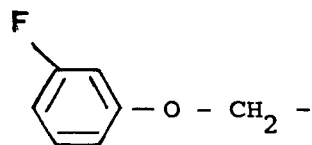


Simple pairs	Augmented pairs	Bonded pairs	Octuplets	Augmented atoms
C * C x6	1 C * C 1 x3	* C * C * x3	C * C * C * C x3	C * C * C x4
C - F x1	1 C * C 2 x2	* C * C * x2 	C * C * C * C x1   F	C * C * C x1   F
C - O x2	2 C * C 2 x1	* C * C * x1 	C * C * C * C x1   O	C * C * C x1   O
	1 C - O 1 x1	- C - O - x1	C * C * C * C x1     F O	C - O - C x1
	2 C - O 1 x1	* C - O - x1 *	C - C - O - C x1	O - C - C x1
	2 C - F x1	* C - F x1 *	C * C - O - C x1 * C	F - C x1
			C * C - F x1 * C	

Continued ...

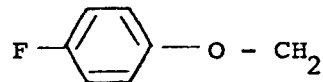
Figure 18 (continued)

Side chain 2



Simple pairs	Augmented pairs	Bonded pairs	Octuplets	Augmented atoms
AS for side-chain 1	1 C * C 1 x2	* C * C * x2	C * C * C * C x2	As for side chain 1
	1 C * C 2 x4	* C * C * x4	C * C * C * C x2	
	2 C - O 1 x1		F	
	1 C - O 1 x1	- C - O - x1	C * C * C * C x2	
	2 C - F x1	* C - O - x1	O	
		* C - F x1	C - C - O - C x1	
		*	C * C - O - C x1	
			*	
			C	
			C * C - F x1	
			*	
			C	

Side chain 3



Simple pairs	Augmented pairs	Bonded pairs	Octuplets	Augmented atoms
AS for side-chains 1 & 2	As for side-chain 2	As for side-chain 2	As for side-chain 2	As for side-chains 1&2

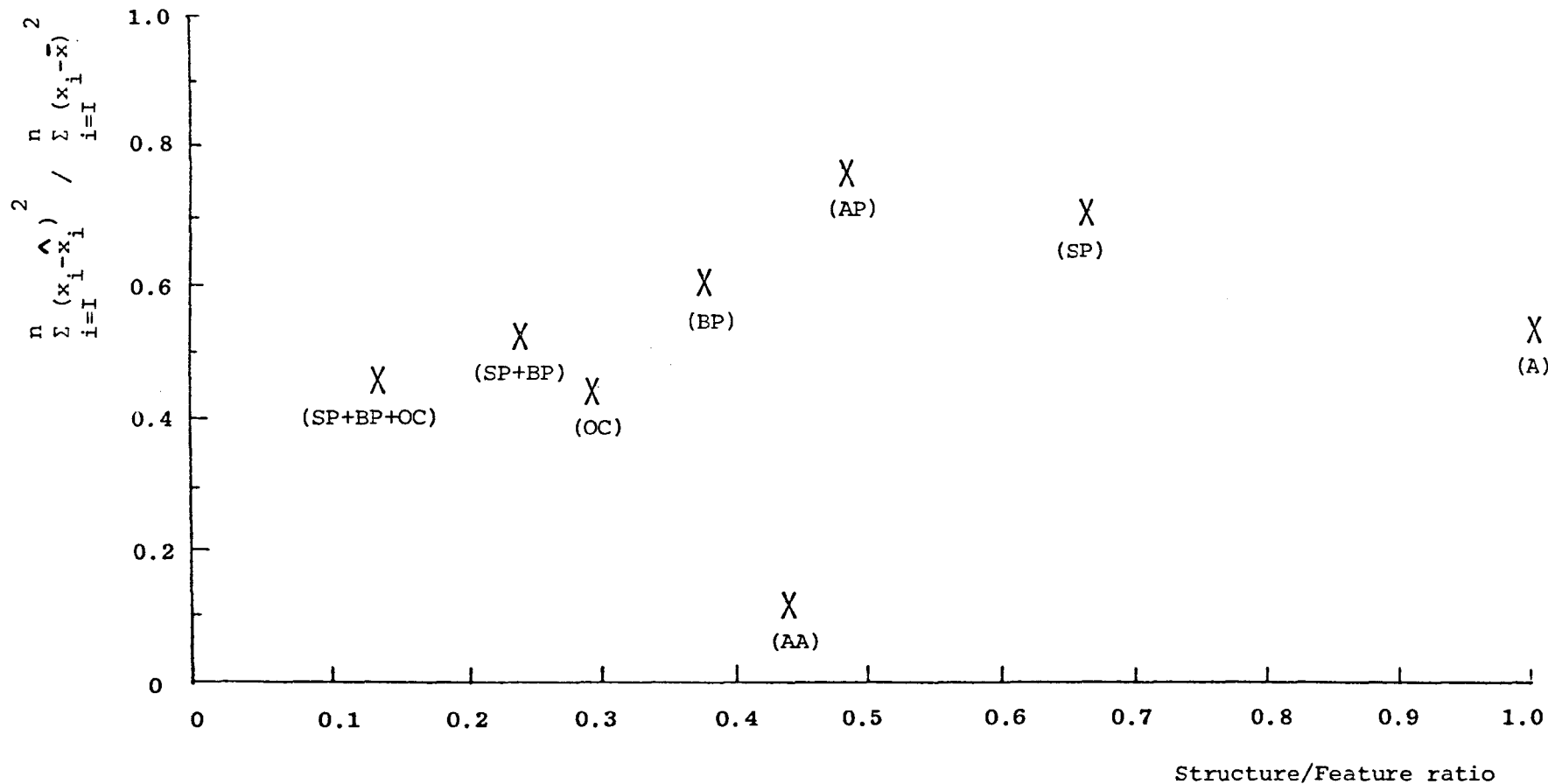


Figure 19 Property (pI) deviations by the nearest neighbour method in 20 amino acids, using a variety of fragment definitions. In this figure and figures 20 to 24 fragments are compared using the structure to feature ratio (see fragment key in notes to figures), and structure comparisons are based on Dice's coefficient, using representations equivalent to structure representation (ii) and additive coding. Quantities  $x_i$ ,  $\hat{x}_i$ ,  $\bar{x}$  and  $n$  are defined in Table 4.

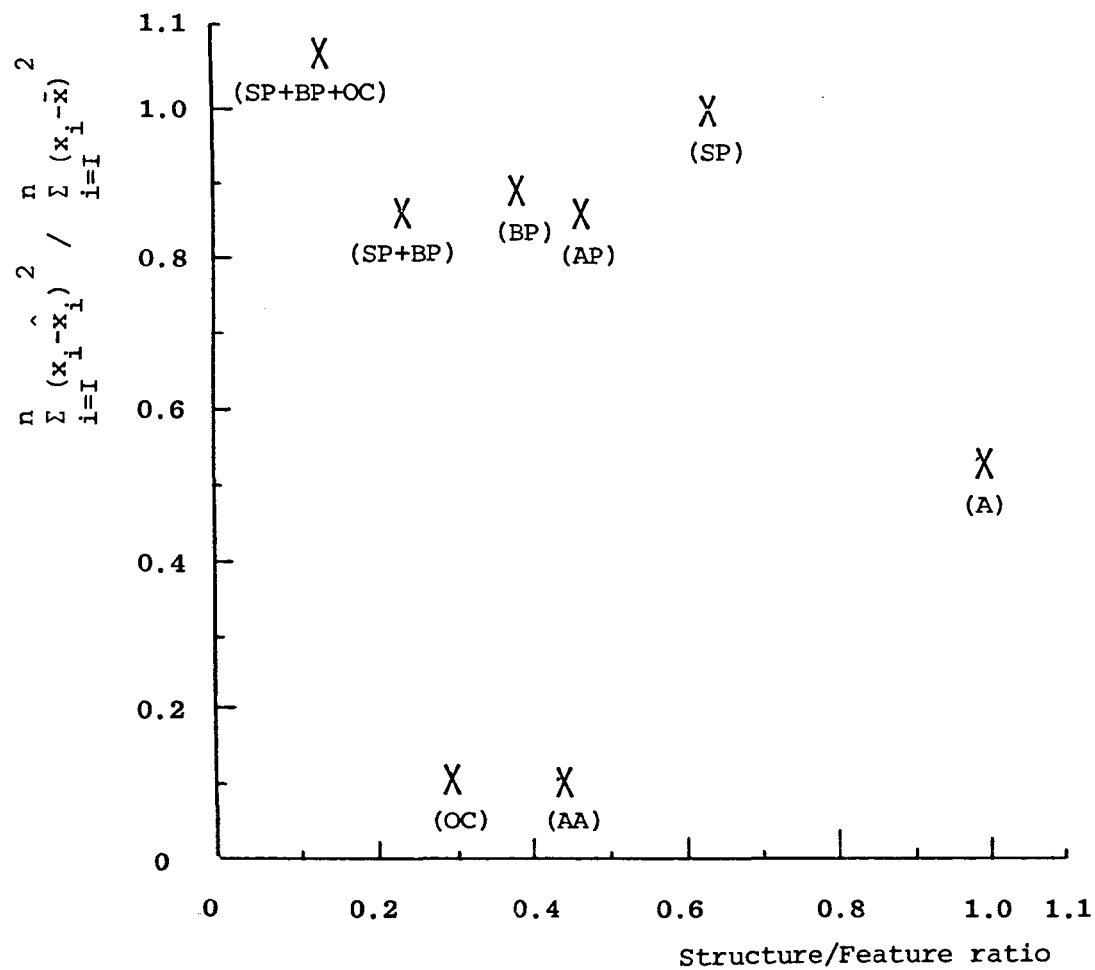


Figure 20 Property (pI) deviations by the classification method (single-link clusters) in 20 amino acids, using a variety of fragment definitions. (see Figure I9)



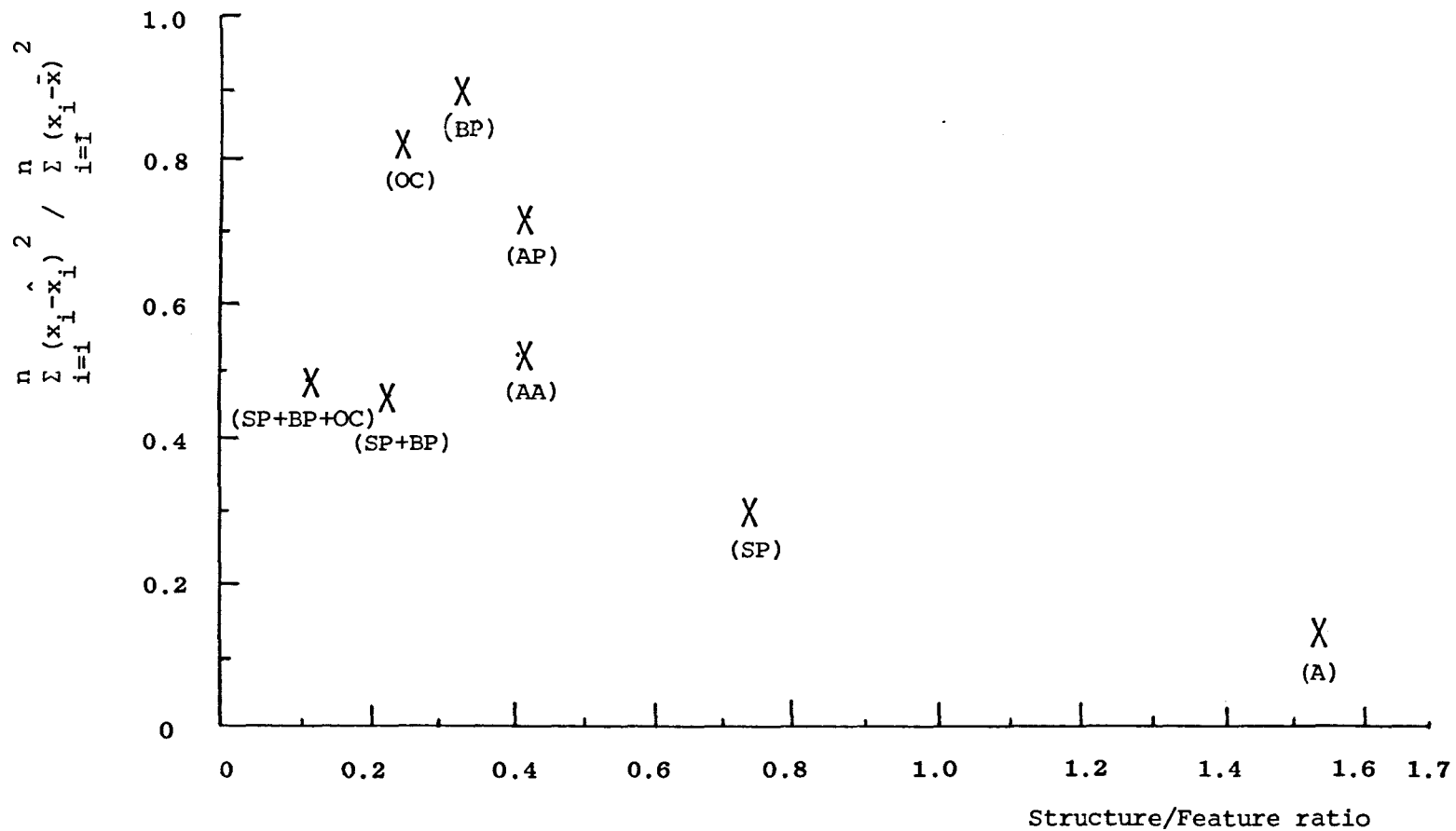


Figure 21 Property (log MBC) deviations by the nearest neighbour method in 39 local anaesthetics, using a variety of fragment definitions. (see Figure I9)

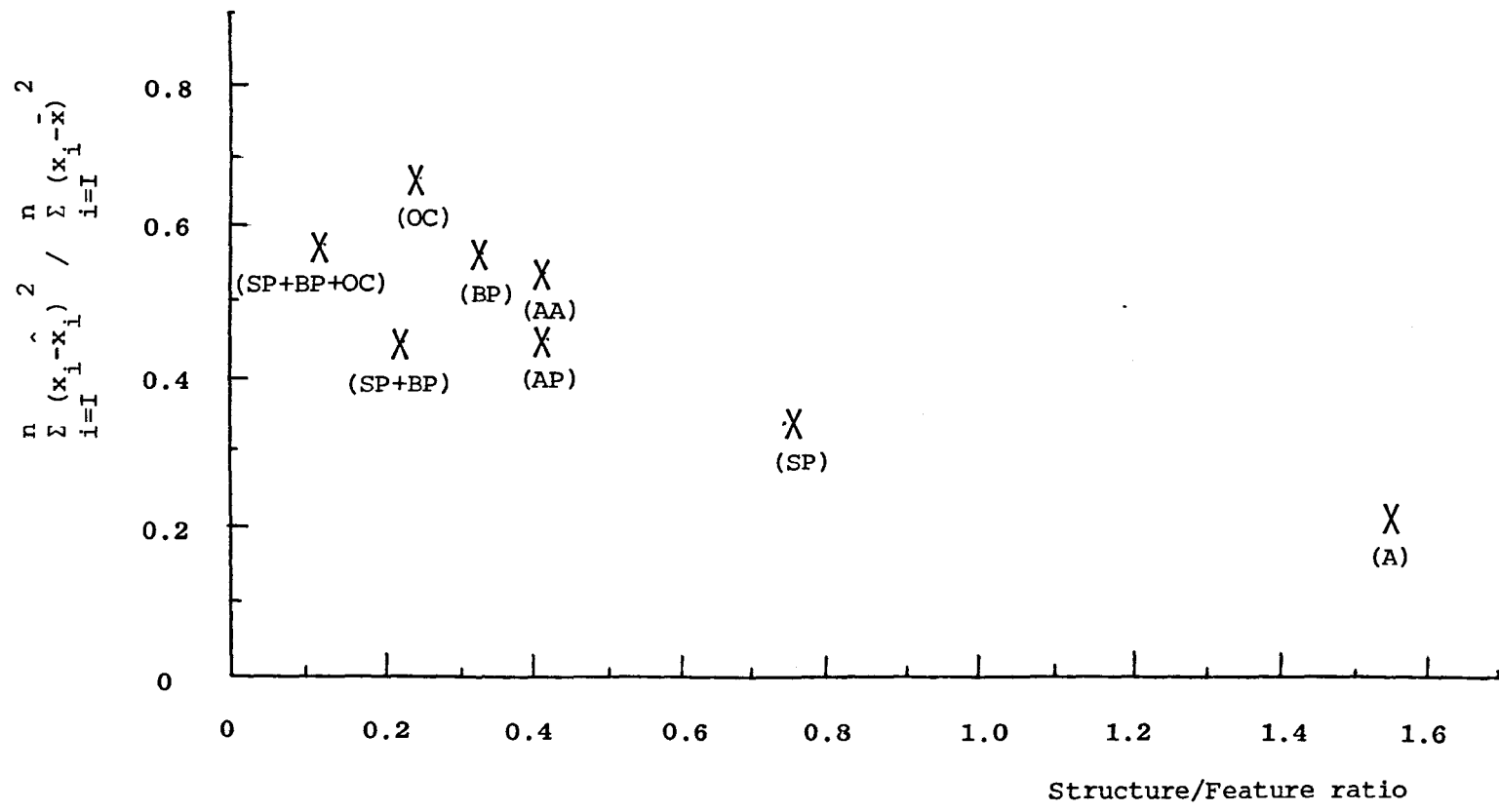


Figure 22 Property (log MBC) deviations by the classification method (single-link clusters) in 39 local anaesthetics, using a variety of fragment definitions. (see Figure I9)

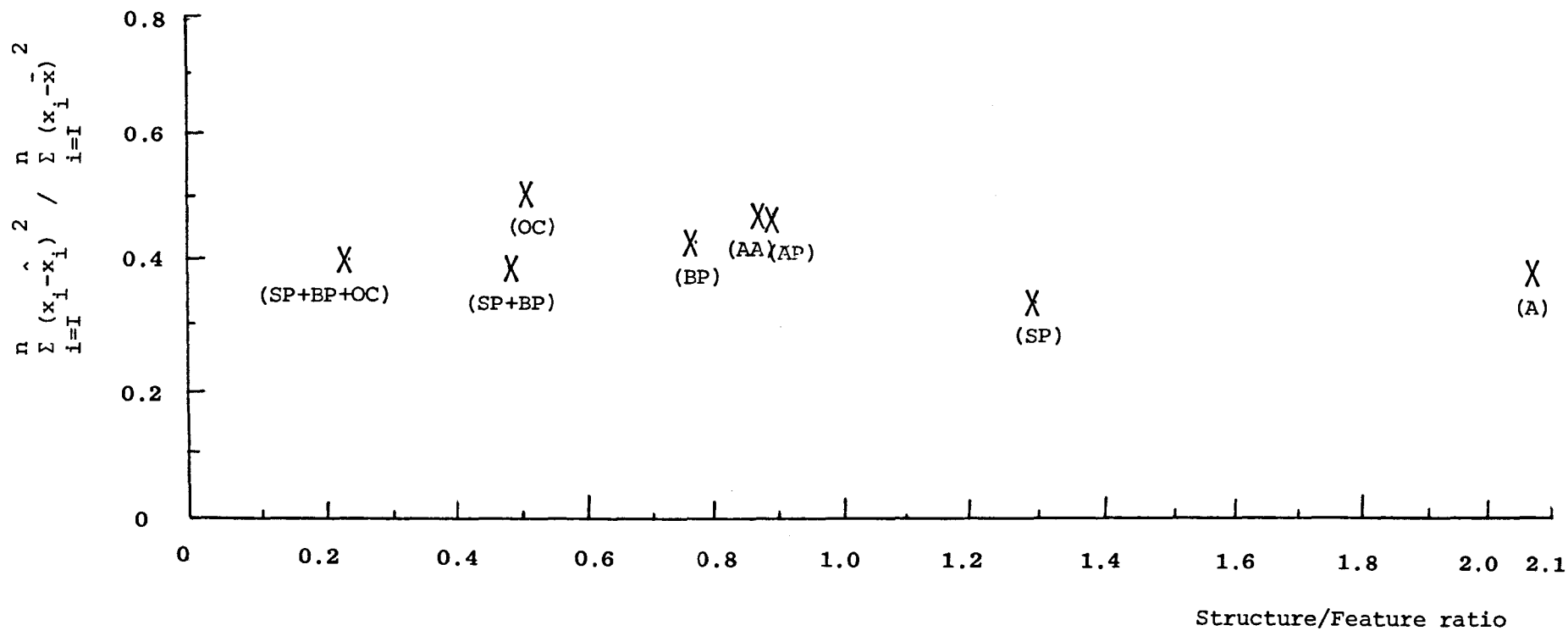


Figure 23 Property (log B/F) deviations by the nearest neighbour method in 79 penicillins, using a variety of fragment definitions. (see Figure I9)

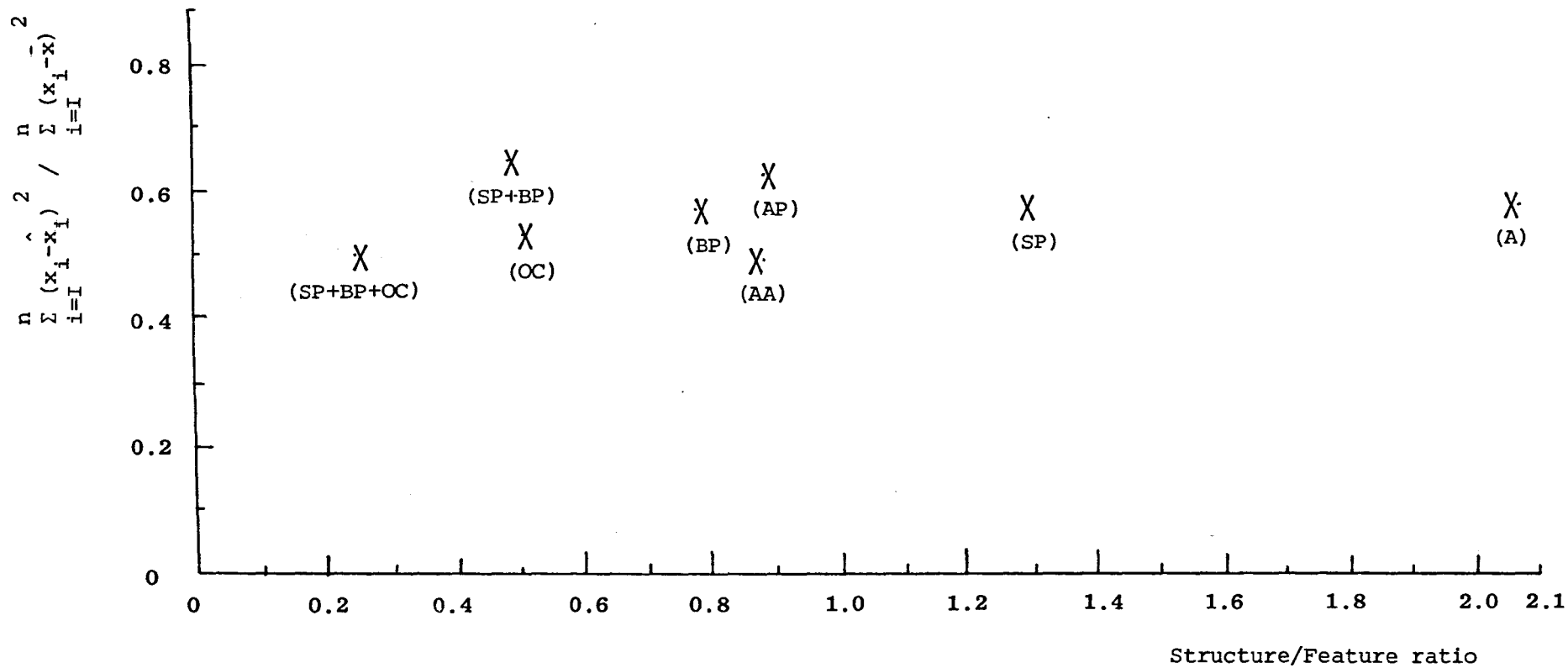


Figure 24 Property (log B/F) deviations by the classification method (single-link clusters) in 79 penicillins, using a variety of fragment definitions. (see Figure I9)

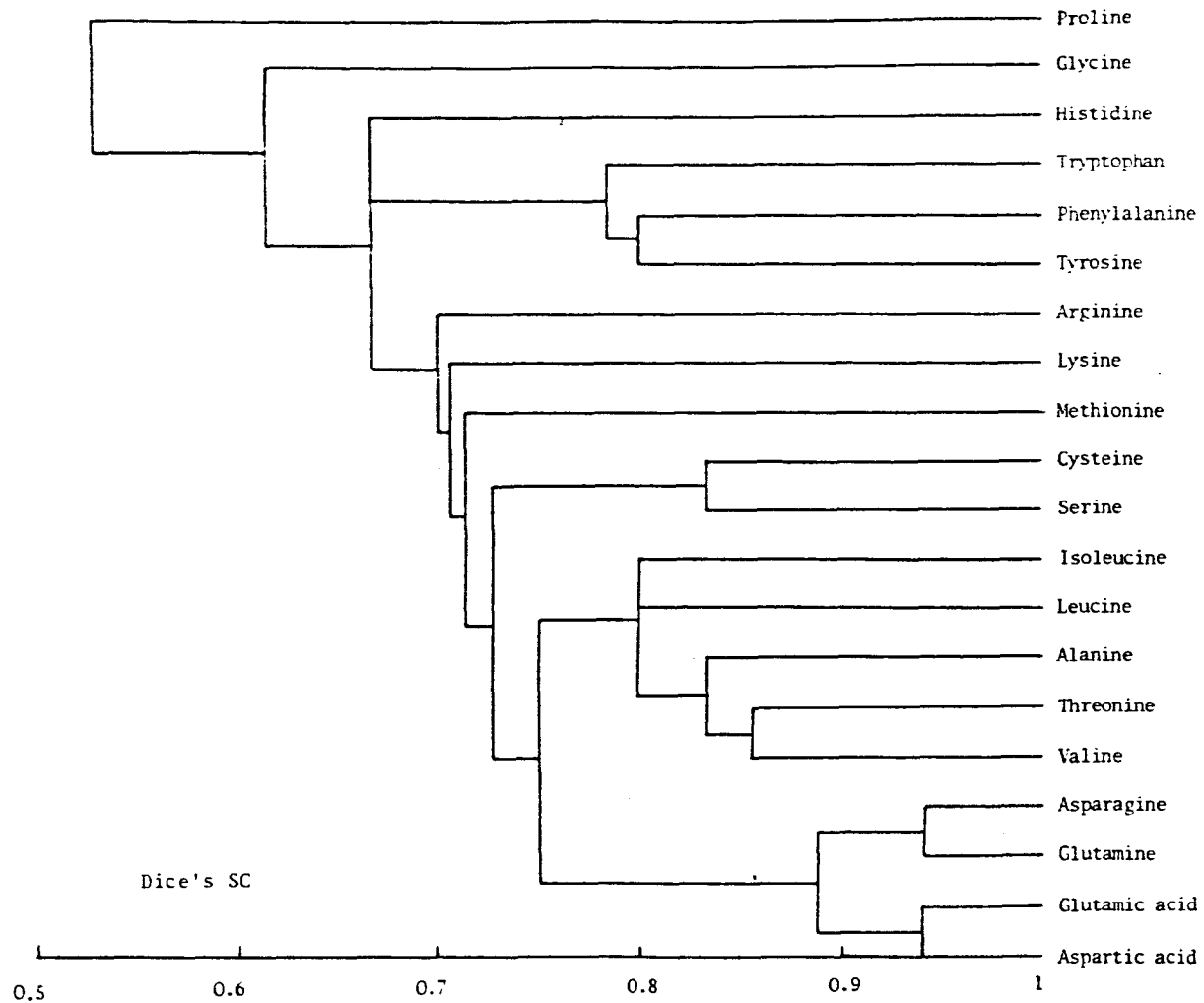


Figure 25 Dendrogram showing the classification obtained for 20 amino acids using structure representations based on augmented pair descriptors. Structure comparisons in this and all following classifications were based on Dice's coefficient, using representations equivalent to structure representation (ii) and additive coding.

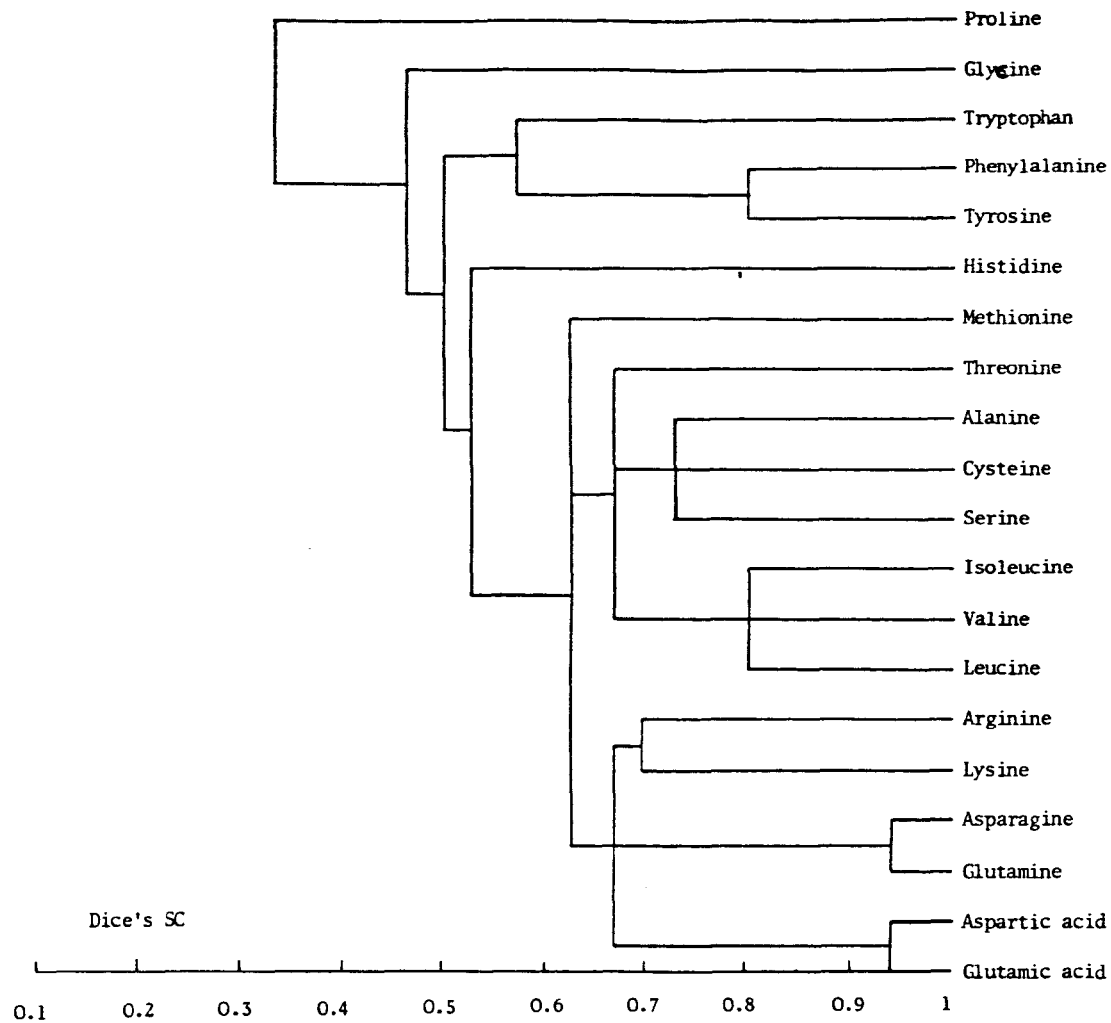


Figure 26 Dendrogram showing the classification obtained for 20 amino acids using Dice's coefficient and octuplet descriptors. (see Figure 25)

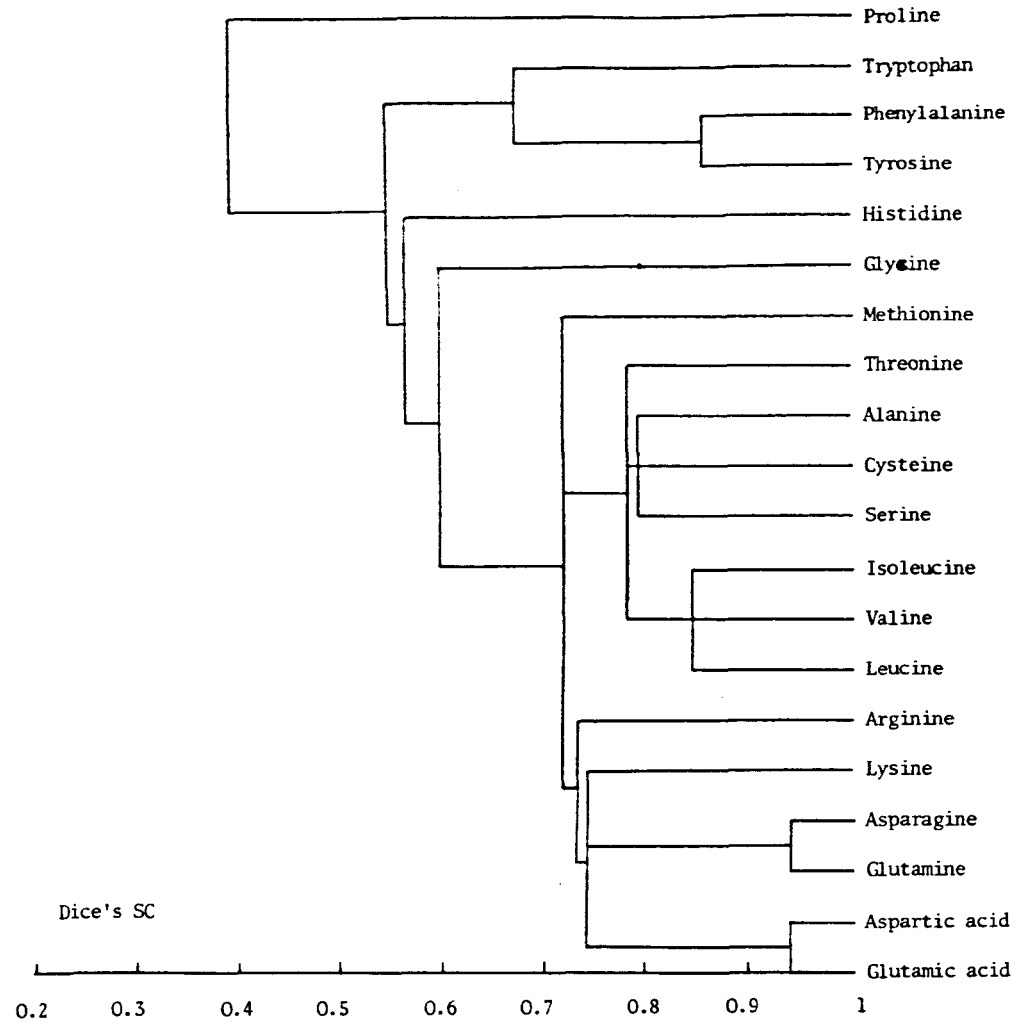


Figure 27 Dendrogram showing the classification obtained for 20 amino acids using Dice's coefficient and simple pair, bonded pair and octuplet descriptors. (see Figure 25)

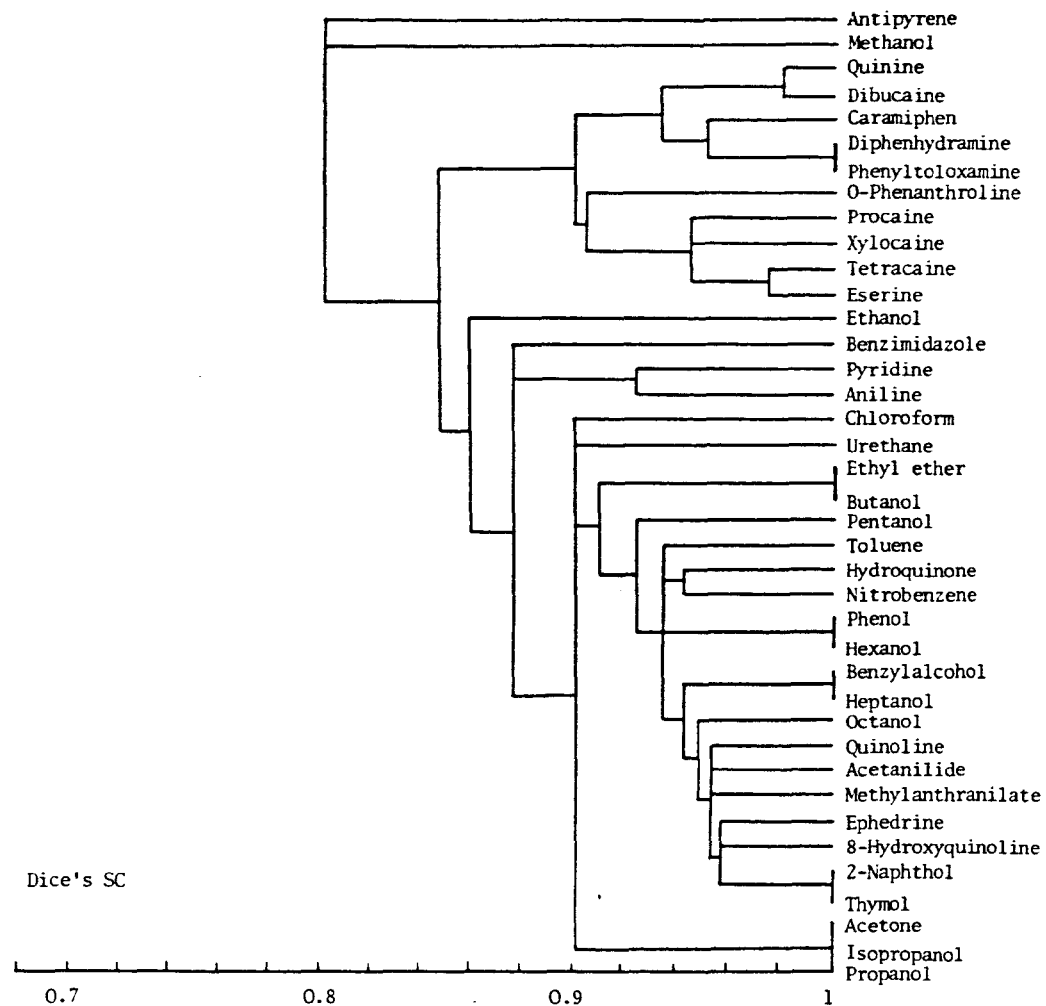


Figure 28 Dendrogram showing the classification obtained for 39 local anaesthetics using Dice's coefficient and atom descriptors. (see Figure 25)



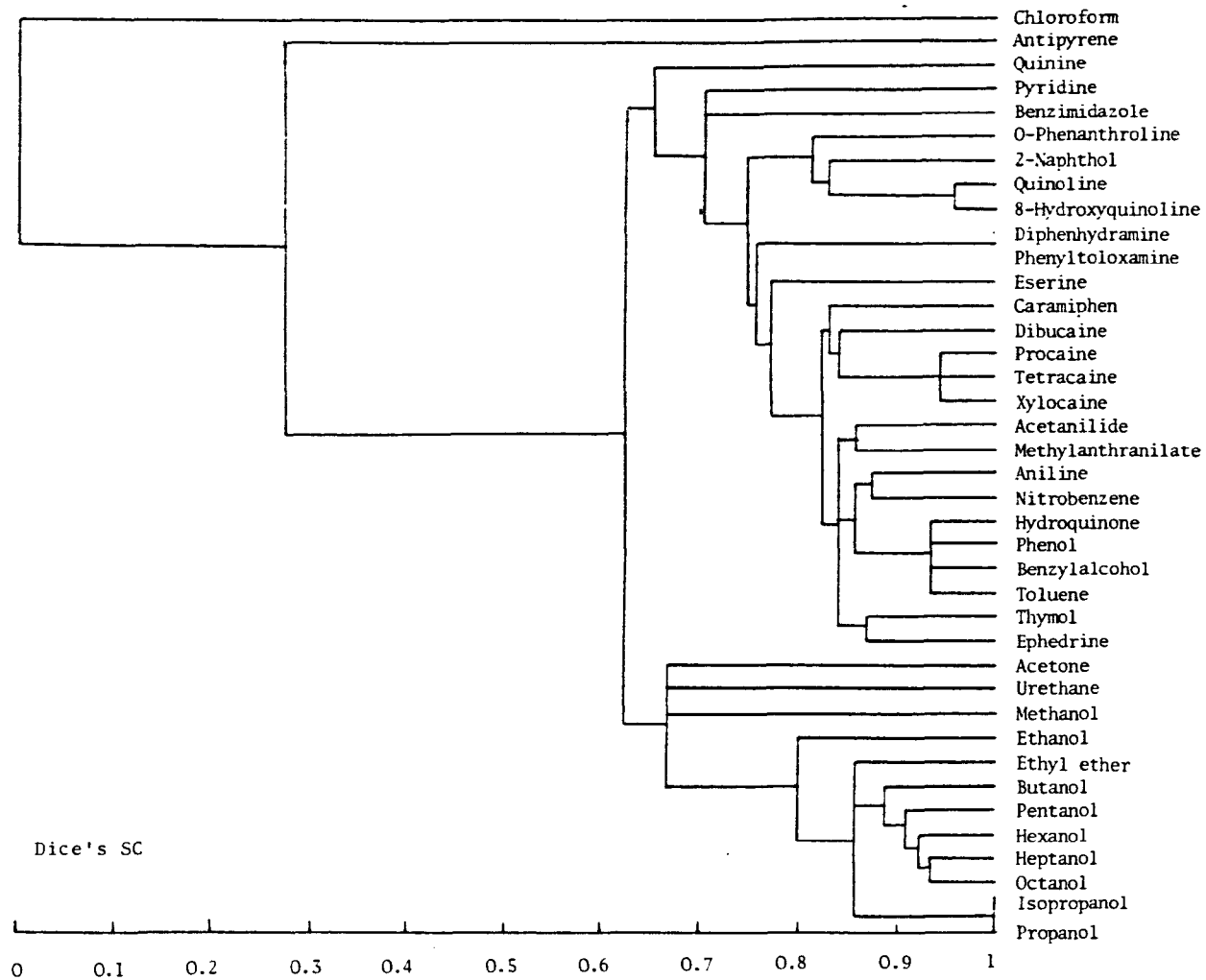
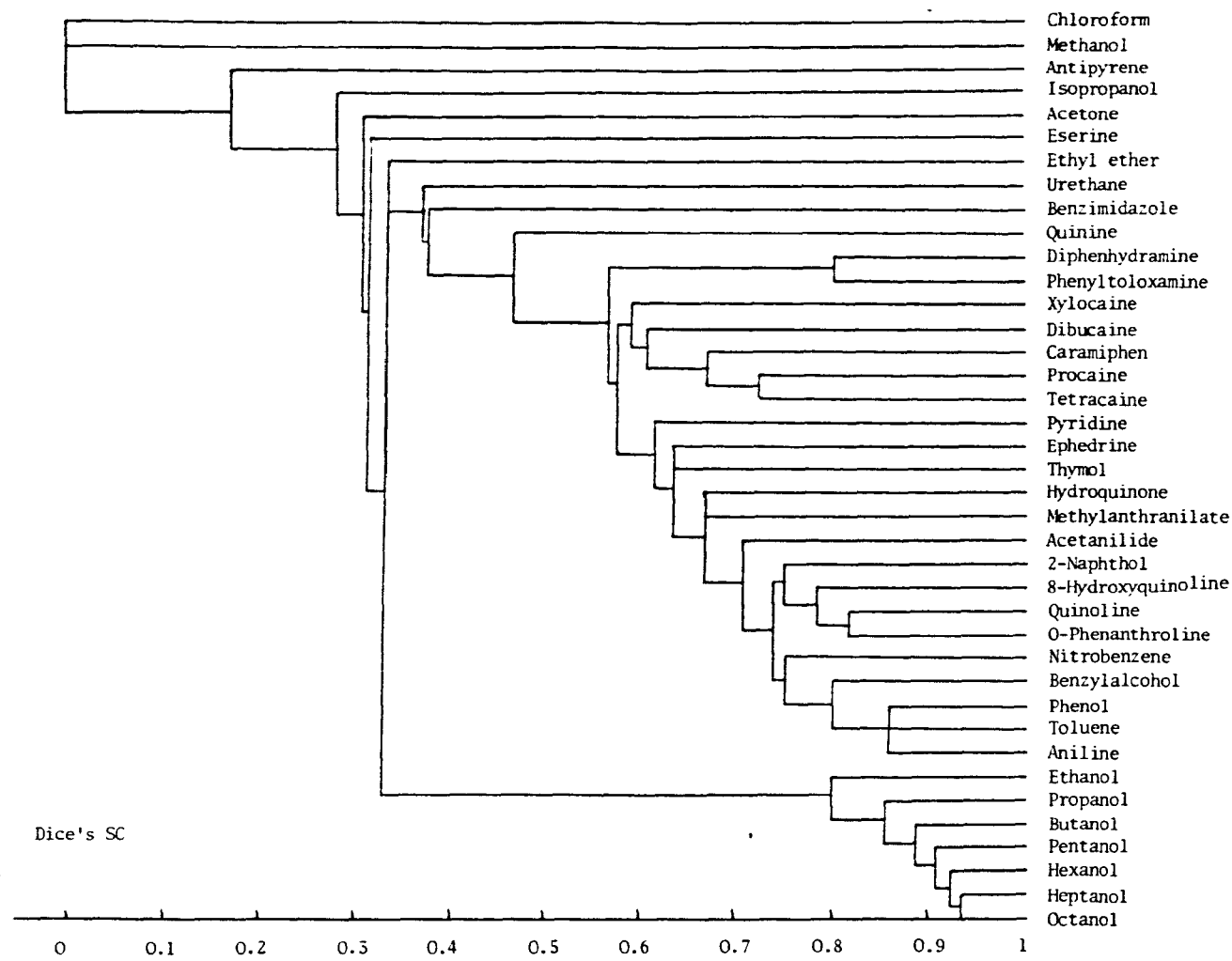


Figure 29 Dendrogram showing the classification obtained for 39 local anaesthetics using Dice's coefficient and simple pair descriptors. (see Figure 25)



**Figure 30** Dendrogram showing the classification obtained for 39 local anaesthetics using Dice's coefficient and bonded pair descriptors . (see Figure 25)

Figure 31 Dendrogram showing the classification obtained for 79 penicillins using Dice's coefficient and augmented atom descriptors (see Figure 25). Structure diagrams are given in Appendix 1.

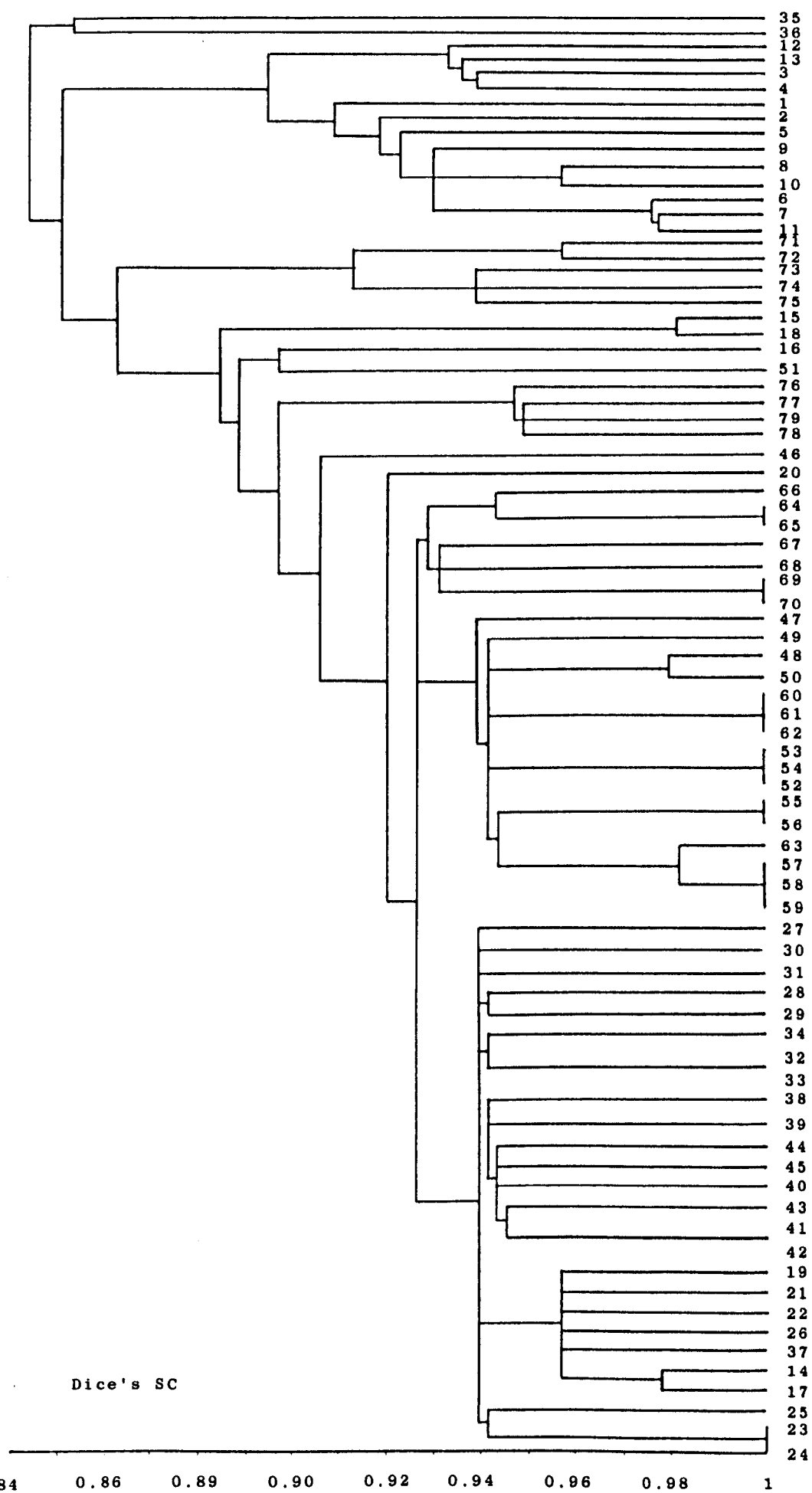


Figure 32 Dendrogram showing the classification obtained for 79 penicillins using Dice's coefficient and simple pair descriptors (see Figure 25). Structure diagrams are given in Appendix 1.

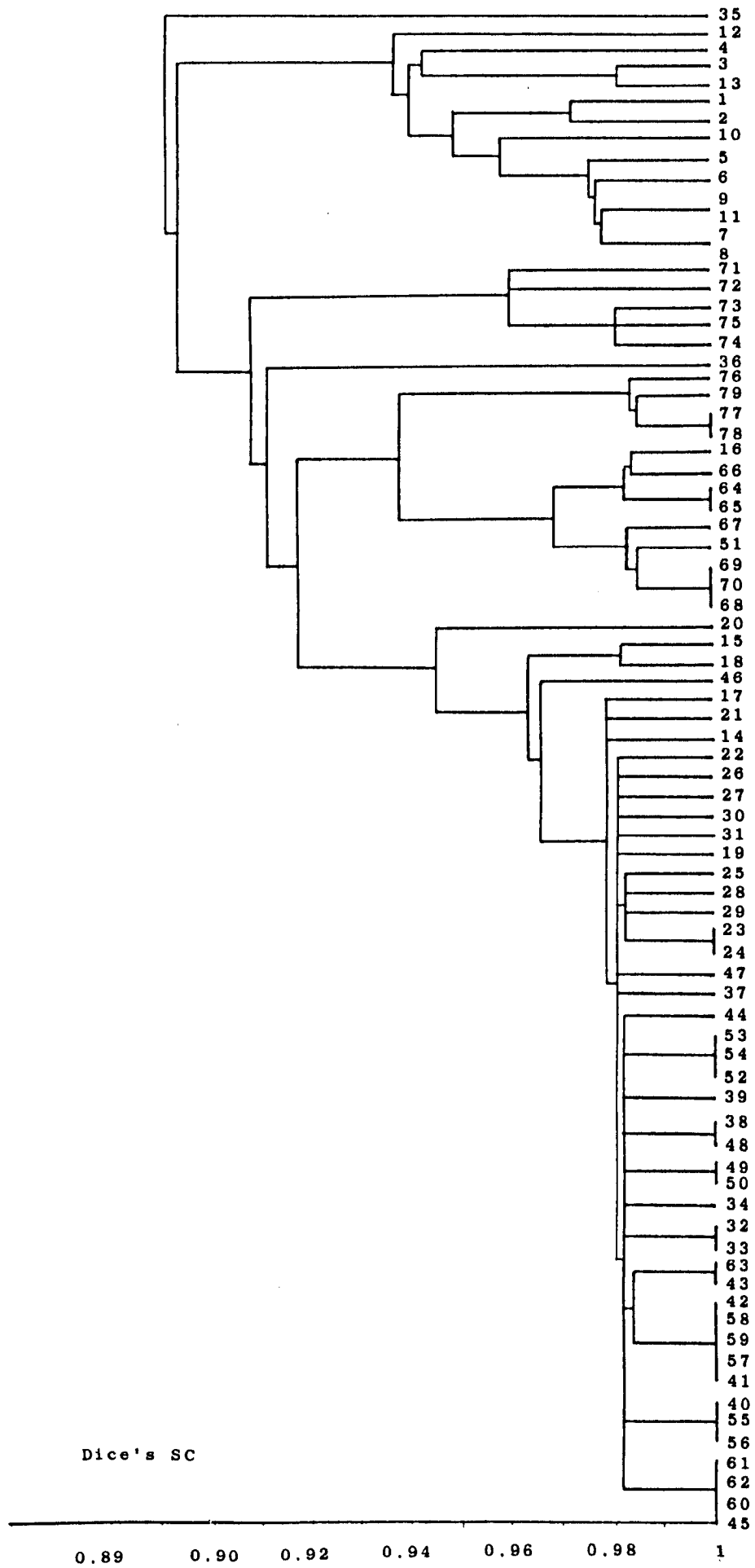
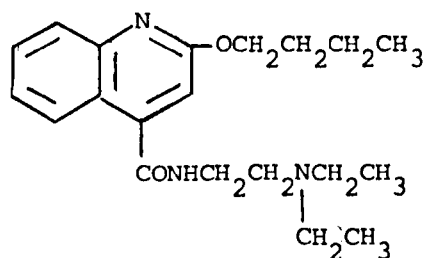




Figure 34 Application of the empirical regression method to predict the log (MBC) value of a local anaesthetic, using the 'hold-one-out' technique. The regression coefficients were obtained at the 10% significance level, using augmented pairs to describe structures

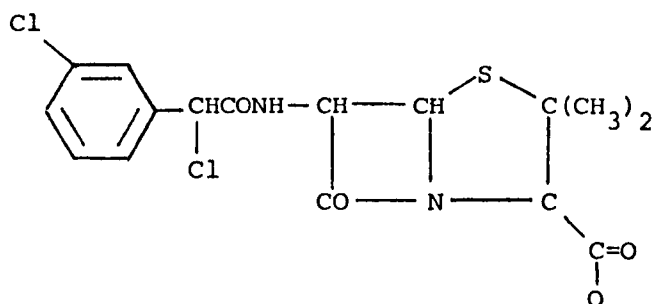


Augmented pair Fragments	Frequency	Regression coefficient x frequency
2C - C2 (chain)	1	-0.750 x 1
1C - N2 (chain)	3	-0.471 x 3
2C * C2	2	-0.527 x 2
1C - C1 (chain)	3	-0.512 x 3
1C * C2	4	-0.298 x 4
1C * C1	3	-0.078 x 3
0C - C1	3	excluded
2C - O1 (chain)	1	excluded
2C = O	1	excluded
1C - O1 (chain)	1	excluded
2C * N1	2	excluded
1C - N1 (chain)	1	excluded
regression output		2.187
	Total	-3.992

ie. The predicted log(MBC) value of this structure is -3.99, observed value -4.20



Figure 35 Application of the empirical regression method to predict the serum binding value of a penicillin, using the 'hold-one-out' technique. The regression coefficients were obtained at the 10% significance level, using augmented atoms to describe structures



Augmented atom fragments <sup>+</sup>	Frequency	Regression coefficient x frequency
$\begin{array}{c} \text{C} \\   \\ \text{C} - \text{C} - \text{Cl} \end{array}$	1	-0.265 x 1
$\begin{array}{c} \text{C} \\ * \\ \text{C} * \text{C} - \text{Cl} \end{array}$	1	-0.041 x 1
Cl - C	2	0.683 x 2
$\begin{array}{c} \text{C} \\ * \\ \text{C} * \text{C} - \text{C} \end{array}$	1	-0.101 x 1
$\begin{array}{c} \text{O} \\    \\ \text{N} - \text{C} - \text{C} \end{array}$	1	-0.249 x 1
HO - C	1	-0.239 x 1
H <sub>3</sub> C - C	2	0.287 x 2
O = C	3	-0.331 x 3
C * C * C	4	0.258 x 4
C - N - C	1	excluded
Regression constant		excluded
Total		1.083

continued..

ie. the predicted log (B/F) value for this structure is 1.084,  
observed value 1.195

+ excluding the fragments of the parent compound which are  
constant in each structure

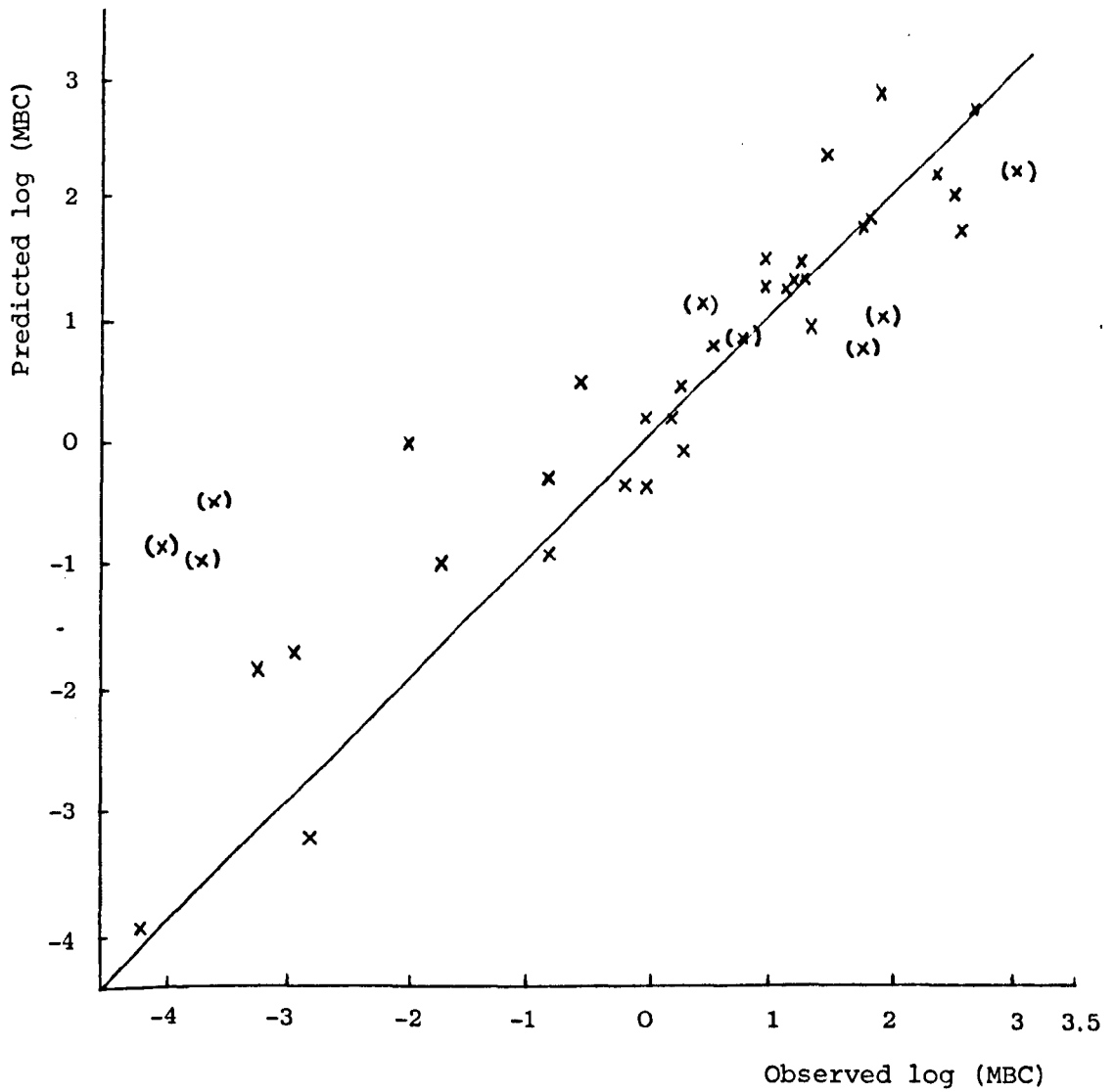


Figure 36 Observed against best 'predicted' log (MBC) values in 39 local anaesthetics by the empirical regression method, using augmented pair descriptors at the 10% significance level. (Predictions based on the 'hold-one-out' technique)

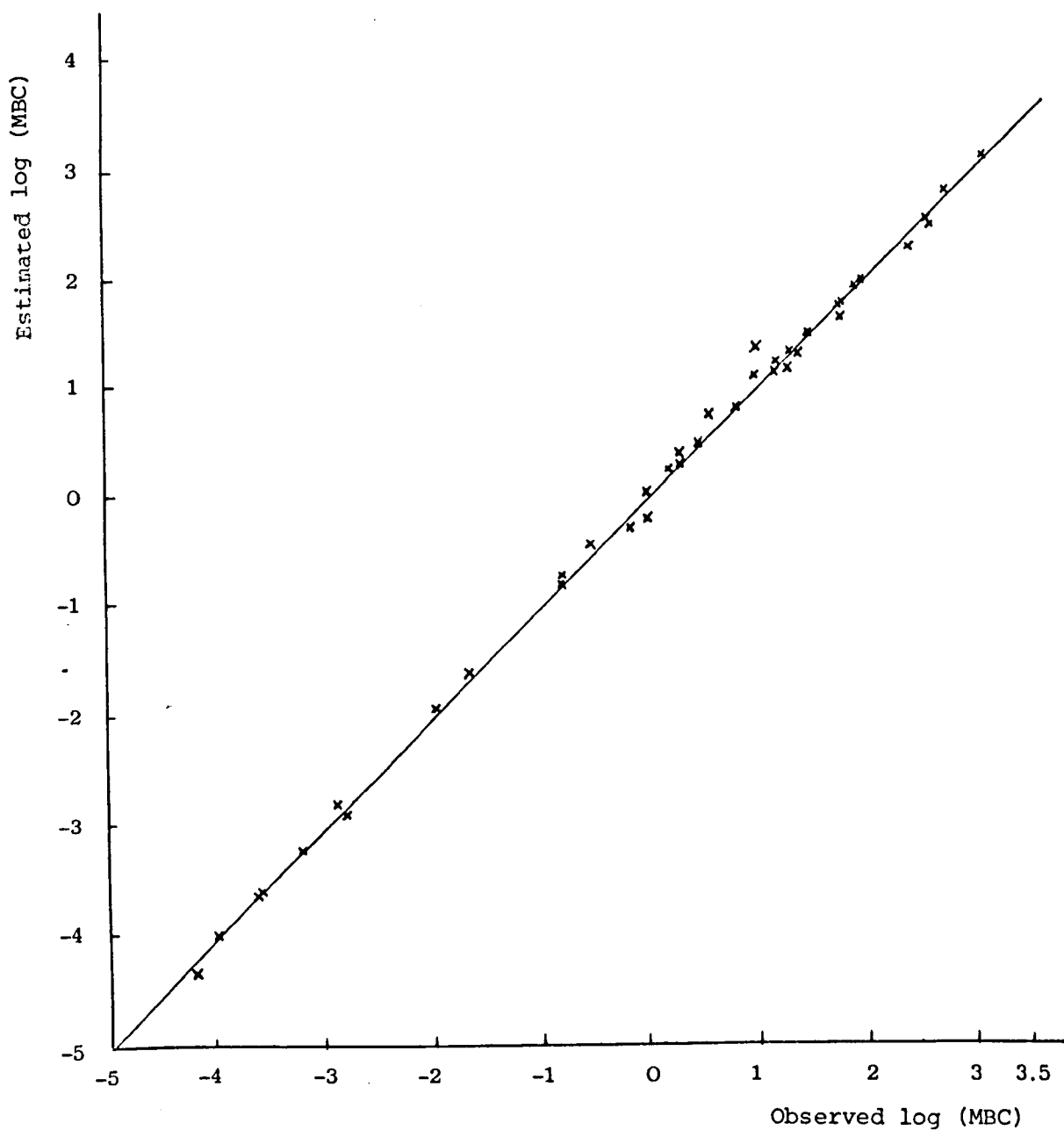


Figure 37 Observed against best estimated log (MBC) values in 39 local anaesthetics by the empirical regression method, using augmented pair descriptors at the 10% significance level. (Estimations from the full structure set).

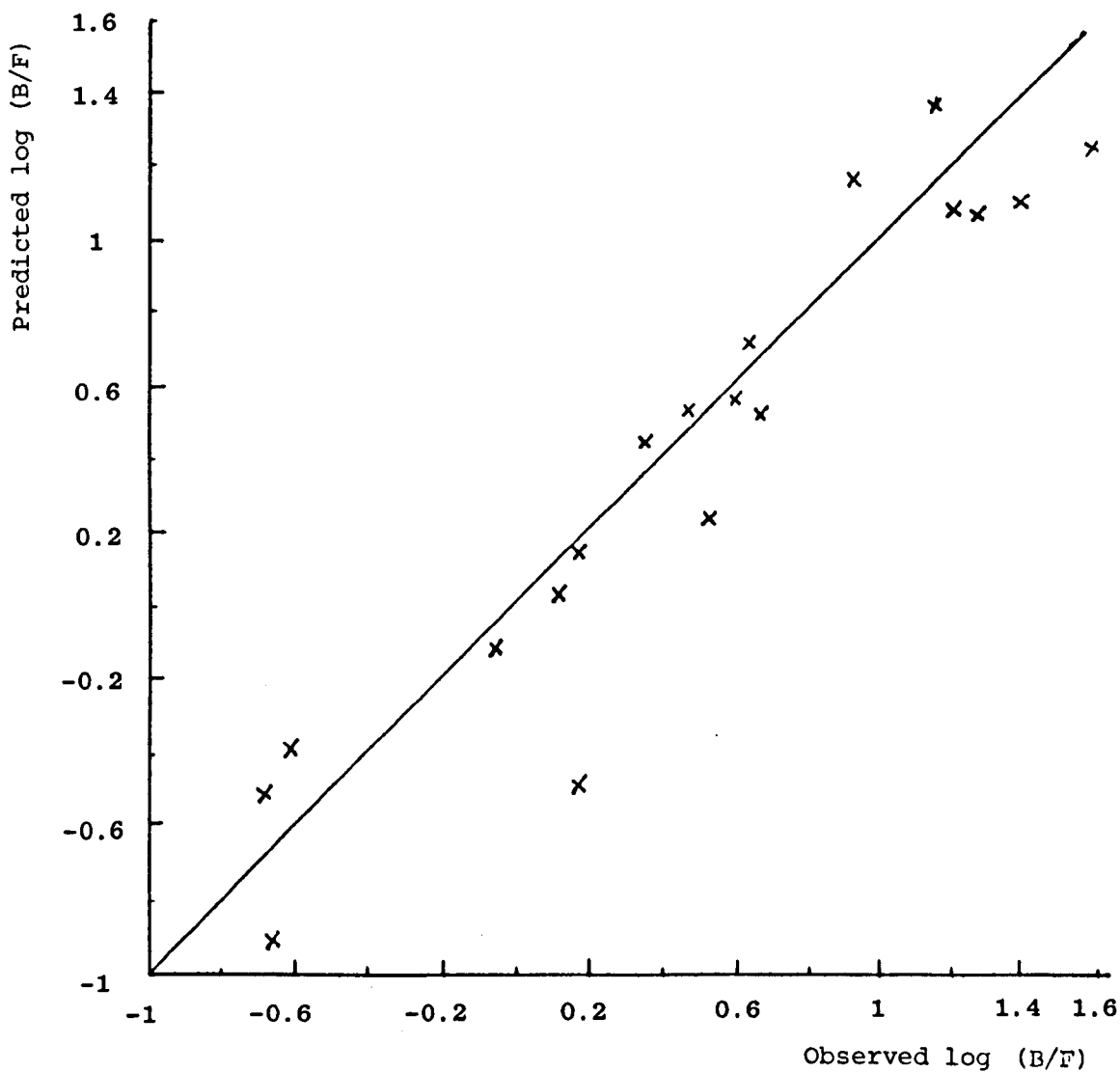


Figure 38 Observed against best 'predicted' log (B/F) values in 20 penicillins by the empirical regression method, using augmented atom descriptors at the 99% significance level. Predictions based on the 'hold-one-out' technique. (Best overall result for this sample)

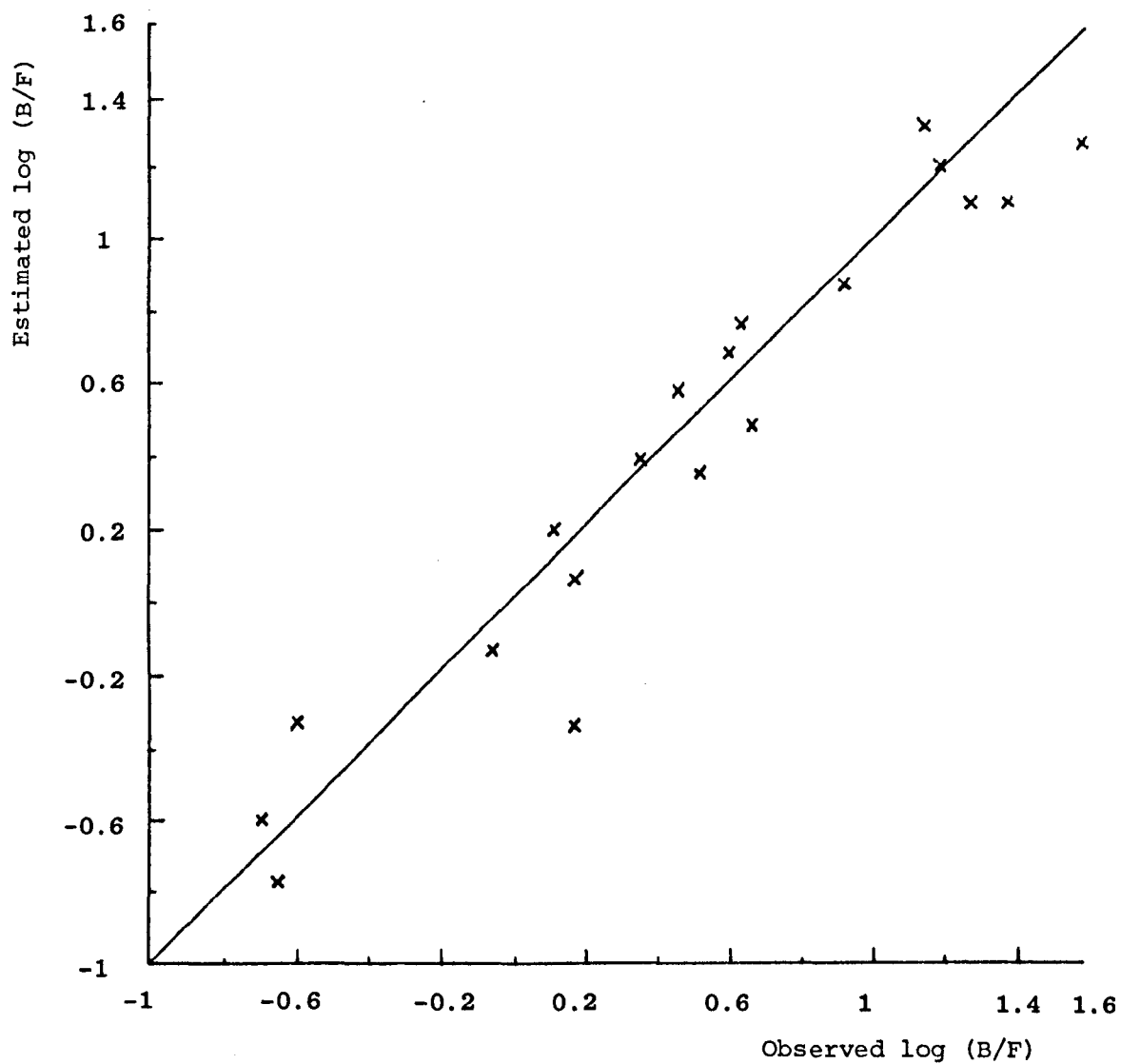


Figure 39 Observed against best estimated  $\log(B/F)$  values in 20 penicillins by the empirical regression method, using octuplet descriptors at the 10% significance level. Estimations from the full structure set.

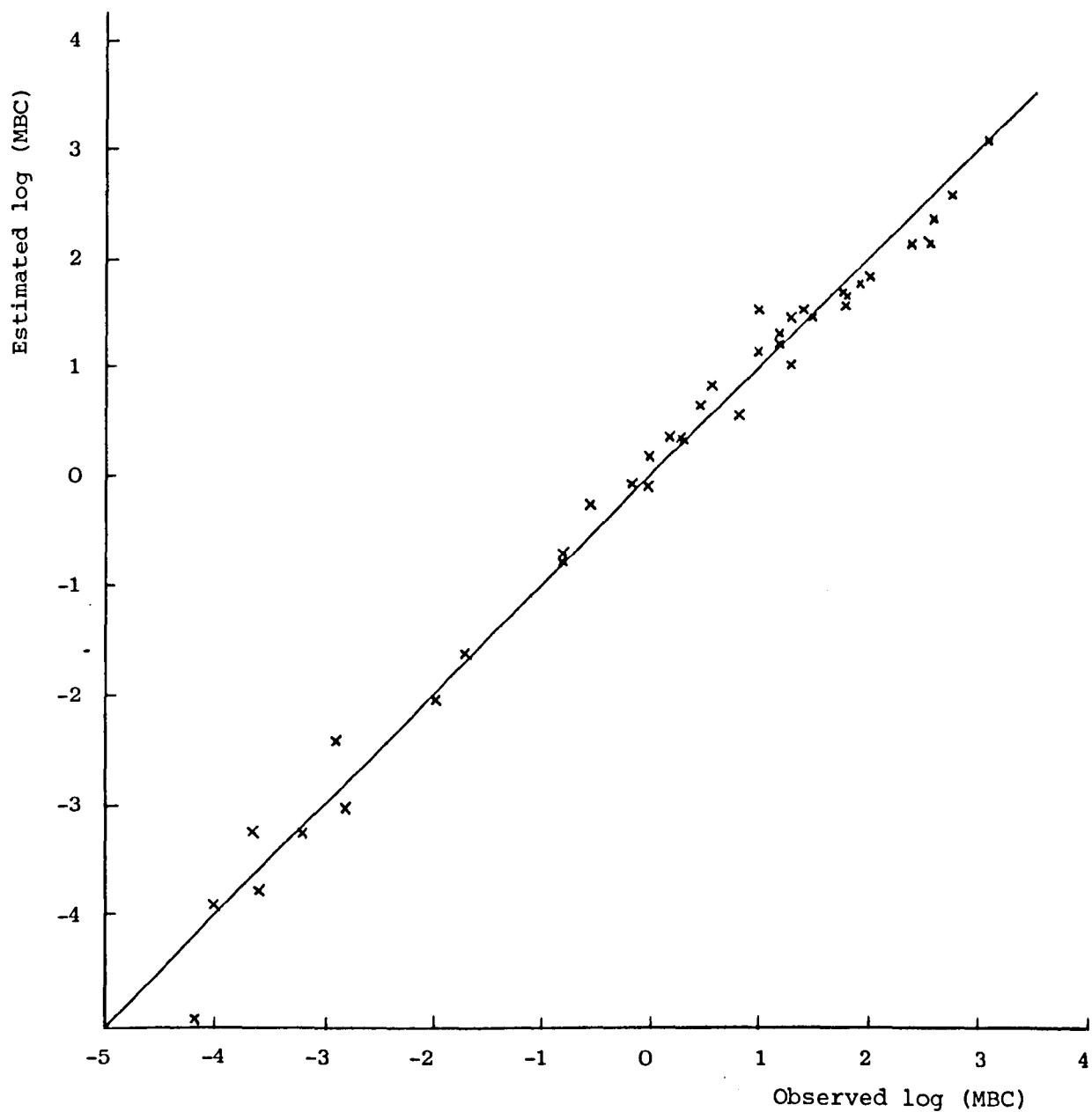


Figure 40 Observed against estimated log (MBC) values in 39 local anaesthetics after Agin et.al <sup>209</sup>. Estimations from the full structure set.

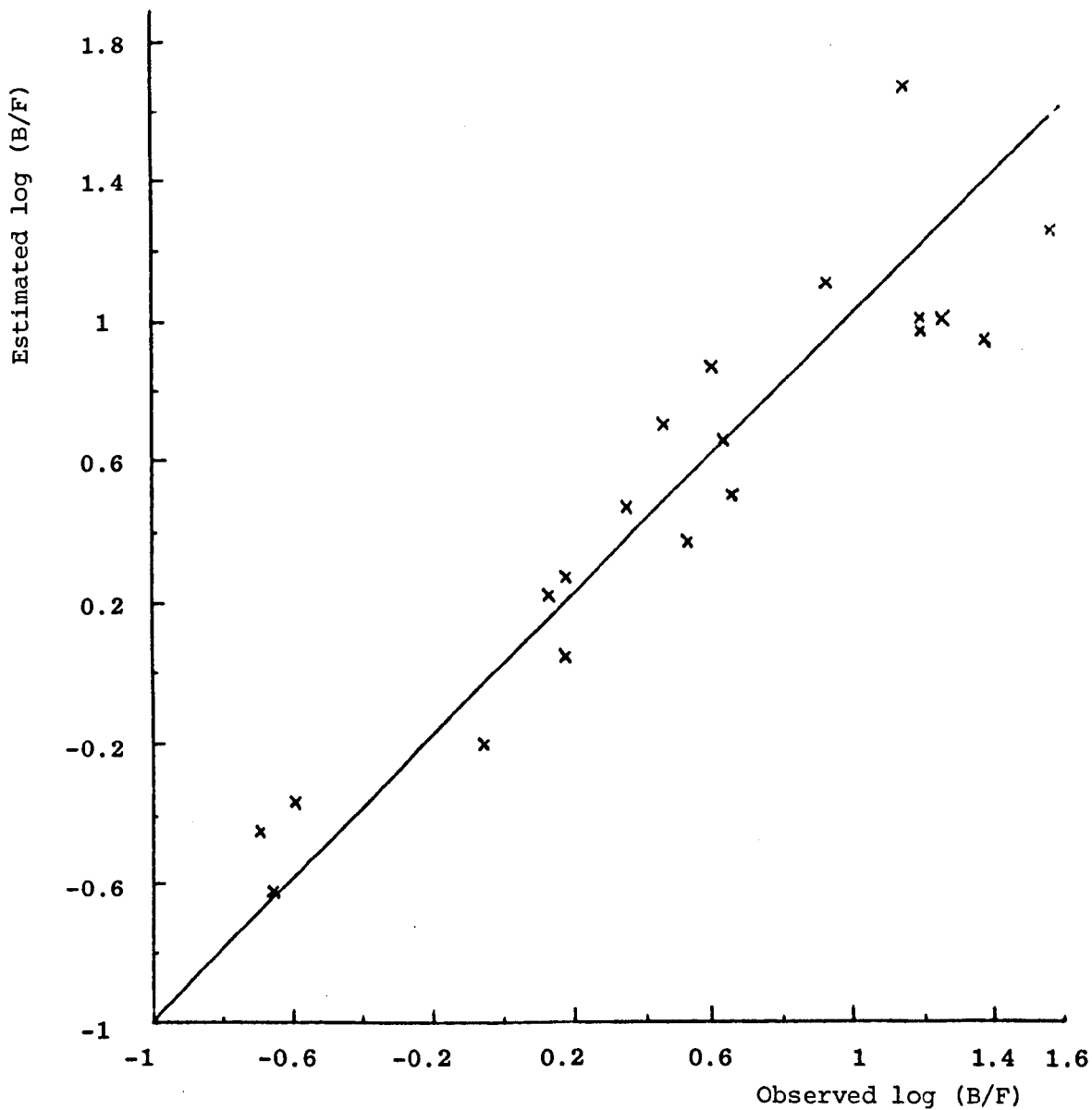


Figure 41 Observed against estimated log (B/F) values in 20 penicillins after Bird and Marshall <sup>210</sup>. Estimations from full structure set.



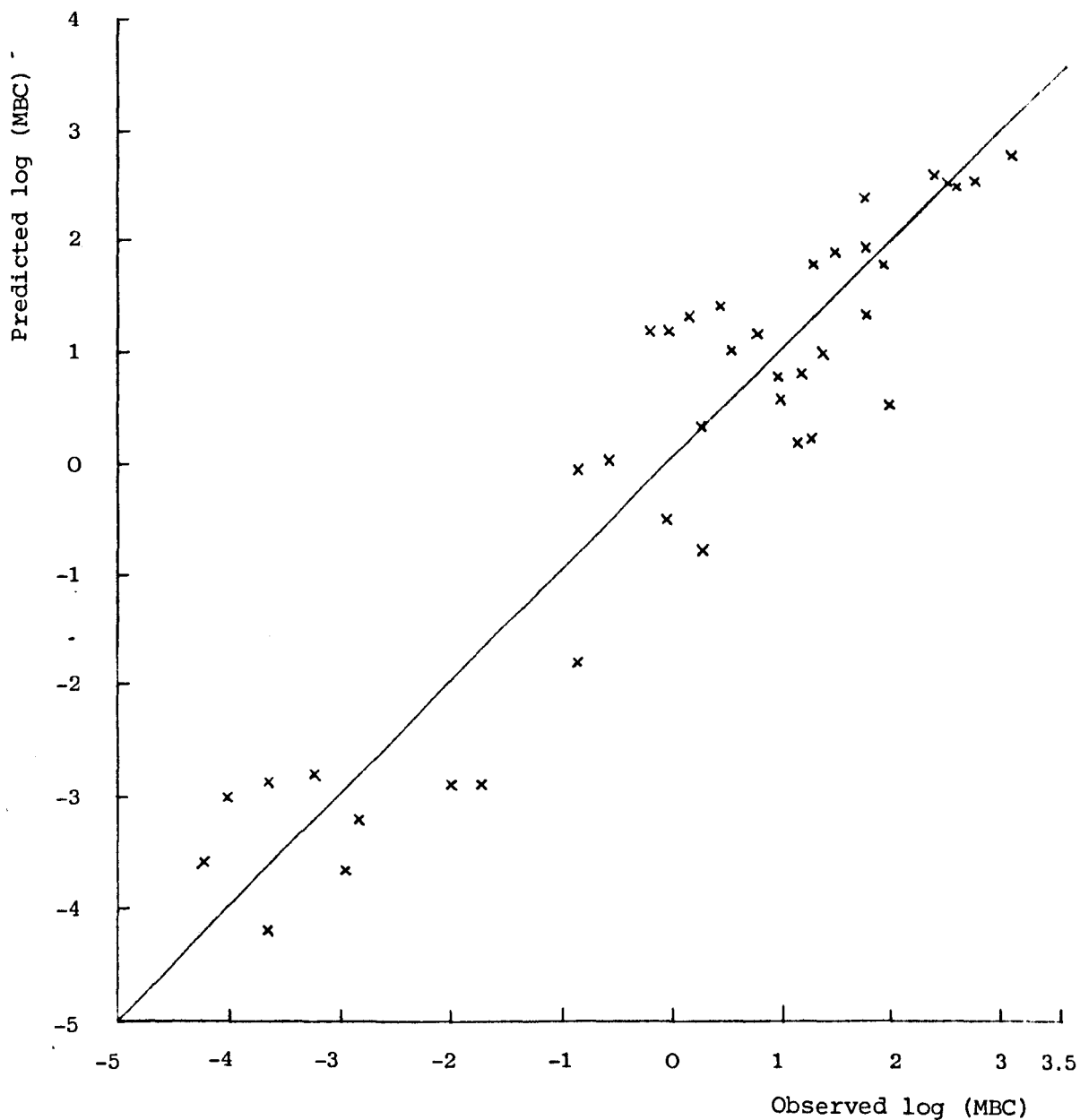


Figure 42 Observed against best 'predicted' log (MBC) values in 39 local anaesthetics by the classification method, using highest associations based on Dice's coefficient (additive coding) and atom descriptors. (Best overall result for this sample)

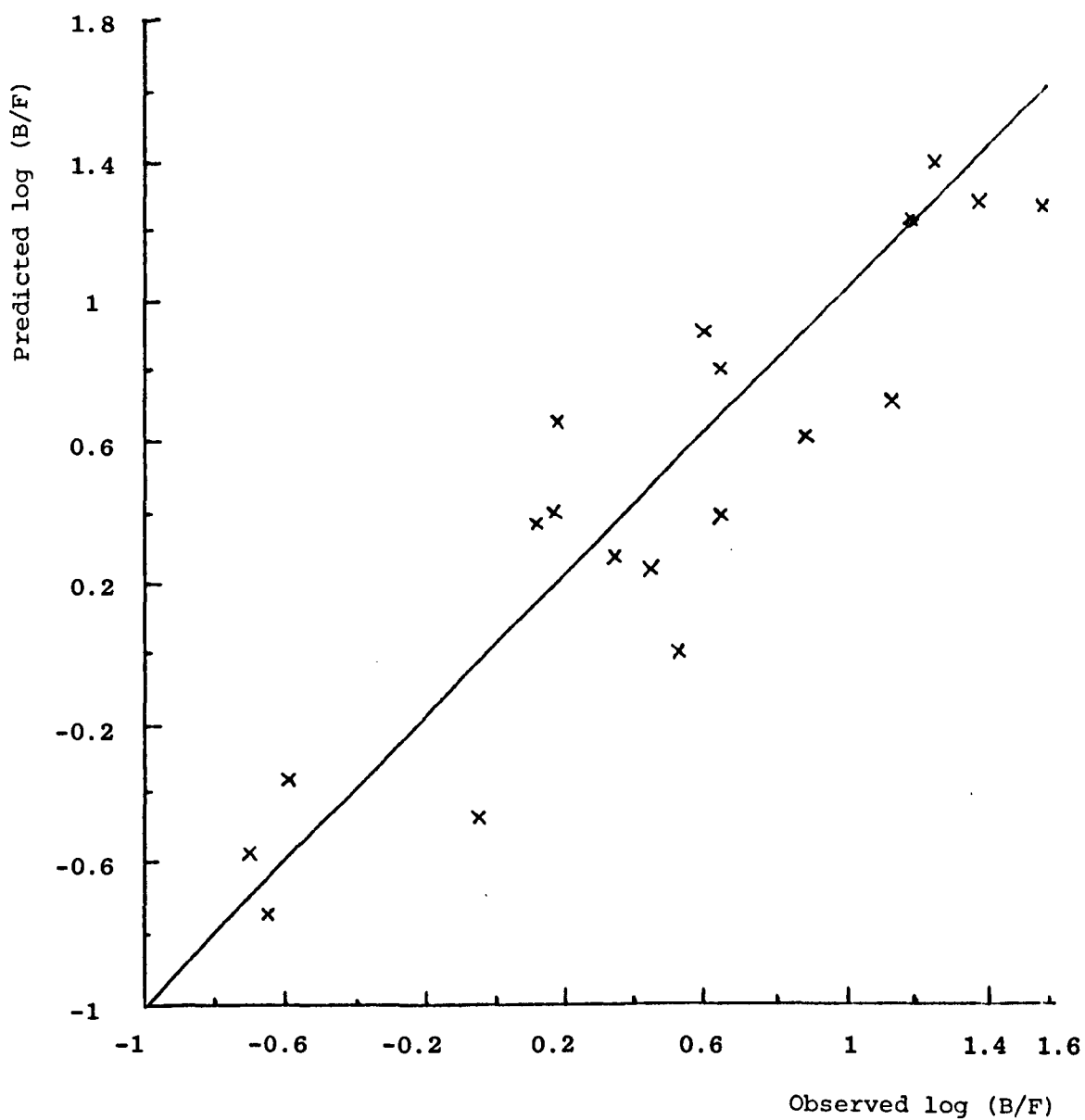
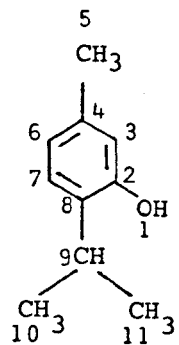


Figure 43 Observed against best 'predicted' log (B/F) values in 20 penicillins by the classification method, using single-link clusters based on the simple distance coefficient and augmented atom descriptors.

Structure



Atom No.	Atom Symbol	Connectivity	Connections							
			Bond Symbol	Atom No.	Bond Symbol	Atom No.	Bond Symbol	Atom No.	Bond Symbol	Atom No.
1	O	01	1	2						
2	C	03	1	1	7	3	7	8		
3	C	02	7	2	7	4				
4	C	03	7	3	1	5	7	6		
5	C	01	1	4						
6	C	02	7	4	7	7				
7	C	02	7	6	7	8				
8	C	03	7	7	7	2	1	9		
9	C	03	1	8	1	10	1	11		
10	C	01	1	9						
11	C	01	1	9						

Figure 44 Example of a redundant connection table record. (Bond types are specified in notes to Figures)

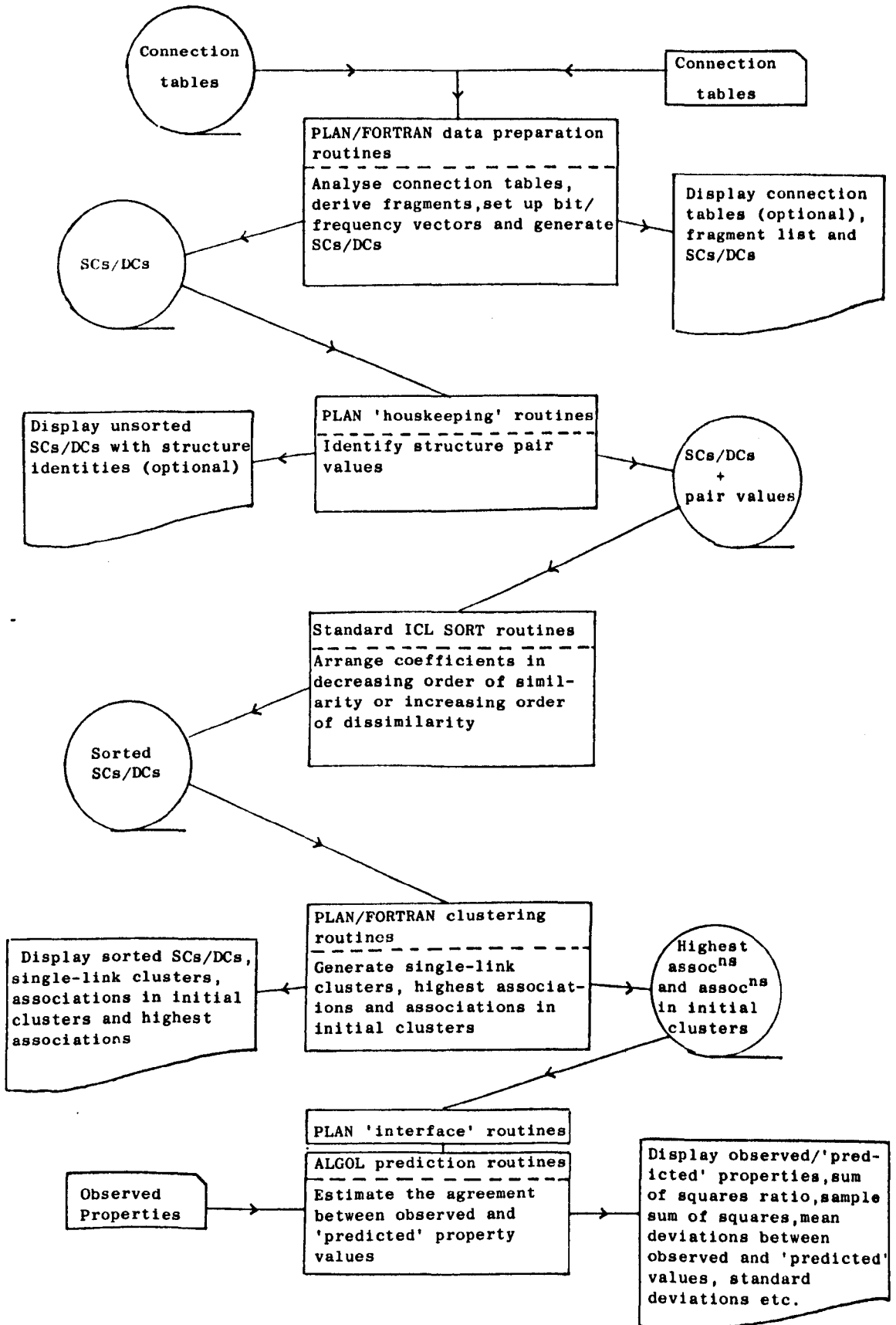


Figure 45 Flowchart of the basic classification procedure.



## PUBLICATIONS

- 1 Adamson, G. W. and Bush, J. A. "A Method for the Automatic Classification of Chemical Structures." Information Storage and Retrieval, 9, 561-568 (1973).
- 2 Adamson, G. W. and Bush, J. A. "Method for Relating the Structure and Properties of Chemical Compounds." Nature, 248, 406-407 (1974).
- 3 Adamson, G. W. and Bush, J. A. "A Comparison of the Performances of some Similarity and Dissimilarity Measures in the Automatic Classification of Chemical Structures." J. Chem. Inf. Comput. Sci., 15 (1) 55-58 (1975).
- 4 Adamson, G. W. and Bush, J. A. "The Evaluation of an Empirical Structure-Activity Relationship for Property Prediction in a Structurally Diverse Group of Local Anaesthetics." J. Chem. Soc. Perkin I, 168-172 (1976).