Computerized Chemical Structure-Handling Techniques in Structure–Activity Studies and Molecular Property Prediction

DAVID BAWDEN

Pfizer Central Research, Sandwich, Kent CT13 9NJ, England

Received May 26, 1982

Applications of computerized structure-handling techniques to studies of the relationships between molecular properties and chemical structure are reviewed. These applications include physicochemical property prediction, structure–activity correlation, large-scale substructural analysis, and structural classification. A comparison of types of structural descriptor is made, with reference to statistical amenability, interpretability, feature selection and reduction, and mixing of descriptor types. Implications for chemical information systems are discussed.

INTRODUCTION

Techniques for automatic handling of chemical structure representations, within computerized chemical information systems, are well-developed for information storage and retrieval.[1,2] Such techniques are being increasingly applied to studies of structure–property relationships, involving a variety of molecular properties and statistical methods. Some aspects of these applications are discussed here. The aim of the author has been to give a comprehensive coverage of all significant applications reported in the literature up to early 1982.

Many of the structure-handling facilities required for such studies are largely the same as for information storage and retrieval (structure input–output etc.) and are not dealt with in detail. A distinctive feature is the automatic analysis of structural representations to identify aspects of molecular structure which may be relevant to the structure–property problem under consideration, and this will receive most attention here. In a few cases, noncomputerized studies will be discussed, where there is a clear potential for computerization or where a useful comparison with computerized techniques may be made.

A generalized view of computer-aided structure–property studies is shown schematically in Figure 1. The computer-readable file of structural representations is analyzed to generate structural features, for which the presence/absence indication or occurrence counts will be variables in subsequent analyses. This process is termed "feature derivation" or "feature perception". "Descriptor generation" is a more general term for the derivation of all variables to be used in an analysis: these may include descriptors other than purely structural, e.g., parameters from molecular orbital calculations, or calculated structural quantities, e.g., topological indexes. Some aspects of the use of mixed descriptor sets are discussed below.

Substructures thus derived may be used in one of three ways.

(i) Property estimation, by summation of substructural contributions to a thermochemical or physicochemical property: The contributions corresponding to the structural features perceived in the molecular structure of interest are extracted automatically from a data base of substructural contributions. No statistical analysis is carried out in this procedure, although the values of the substructural contributions may have been obtained initially from such an analysis. Such procedures amount to an automation of the well-known additivity schemes for thermochemical and physicochemical properties.[3] They may be particularly important in calculating molecular properties rapidly for large groups of structures, e.g., thermodynamic parameters for computer-aided synthesis planning[4] or lipophilicity parameters for physicochemical property–biological activity correlation.[5] If the structural features perceived correspond to substituents on a common parent structure, such a procedure would be used in conjunction with a data base of substituent constants as an aid to the Hansch form of linear free energy relationship studies.

(ii) Structure–property correlation, with the occurrence (or presence/absence indication) of features within molecular structures used as variables in statistical analyses for correlation of structure with physicochemical properties or biological activities: Dependent upon the type of features used, and the statistical analysis employed, this computer -aided procedure may correspond to established techniques of quantitative structure–activity relationships (QSAR), e.g., Free–Wilson analysis,[6] or may be an entirely new departure, e.g., the substructural analysis methodology, first described by Cramer et al.[7]

(iii) Structural classification, in the widest sense, involving some assessment of the similarities within a set of structures, based on structural features: Although molecular property data is not directly involved in such a procedure, the classification obtained may subsequently be applied in SAR studies, qualitative or quantitative.

Areas such as computer-aided synthetic planning[8] and computerized elucidation of reaction mechanisms[9] will not be discussed per se, although certain relevant aspects of these studies will be noted. Similarly, the use of substructures in spectral simulation and interpretation (see, for example, ref 10–12) are not discussed here.

One constant problem in this area is that it is frequently unclear whether the molecular property under investigation is affected by the whole structure (as in the case of a bulk physical property) or only by a constituent substructure (as in the case of a specific biological activity). In some cases both factors may be involved, as in the case of a series of compounds with specific pharmacological activity (invoked by a substructure) modified by the compounds' distribution properties (such as the partition coefficient; bulk properties of the total structure). It is therefore necessary to be able to identify and analyze either or both of these structural effects.

OVERVIEW OF TYPES OF STRUCTURAL FEATURES

The types of structural features used in structure–property studies are now outlined briefly. It should be noted that the subdivisions used, although covenient, are far from precise delimitations, as will be noted below.

(i) Simple Features. The simplest types of structural feature which may be used are counts of the most basic structural units present: total number of atoms, bonds, or rings, occurrence of particular atoms, multiple bonds, etc. Descriptors of this sort are in general too crude to be used alone in structure–property studies, since they are insufficiently discriminating between structures (other than in a poorly structured data set), and their use can make interpretation difficult. A study in which features of this type were used alone in correlation of

0095-2338/83/1623-0014$01.50/0

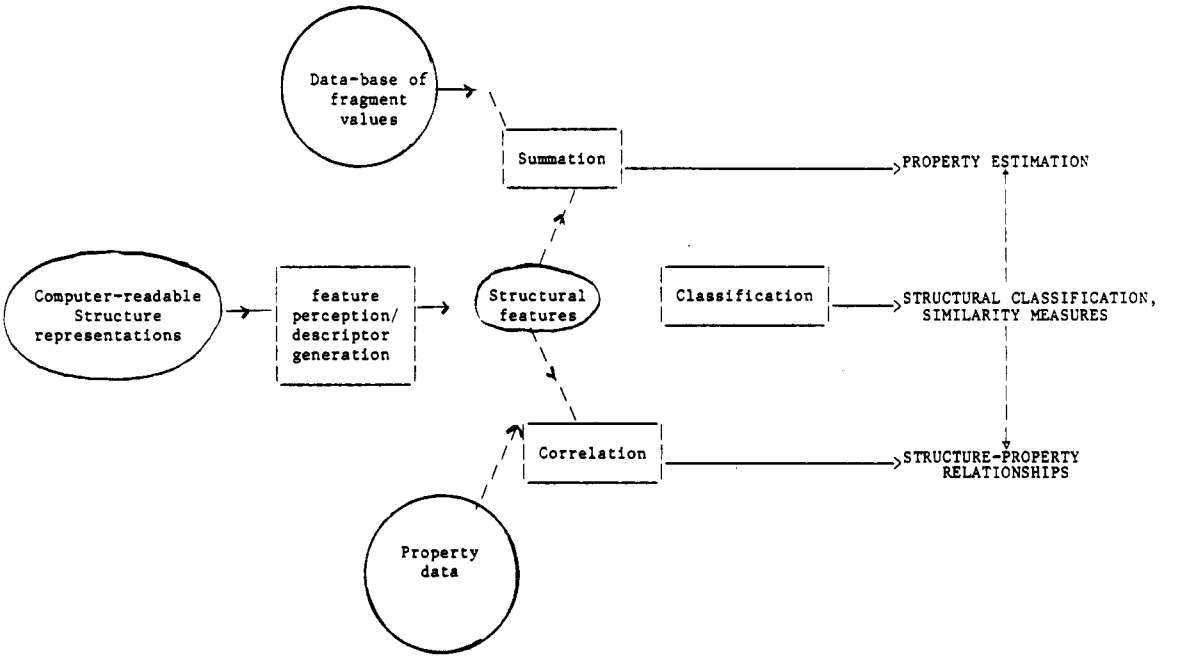

Figure 1. Overview of substructural analysis procedures. The solid arrows indicate a direct relation and the dashed arrows an indirect one.

antitumor activity[13] was criticized on these grounds[14] as is discussed further later.

Features of this sort are frequently used as part of a set of mixed descriptors. In combination with more detailed substructures they have been used in multiple regression and discriminant analyses of acute toxicity, carcinogenicity, and mutagenicity,[15-18] regression analyses of odor intensity,[19] and substructural analyses of antiarthritic activity.[7] They have also been used as variables in ISODATA clustering of structure studies[17] and in conjunction with several other types of descriptor, in pattern-recognition studies of various biological activities.[20-24] It may also be noted that the simple feature arises naturally as the lowest levels of the atom- and bond-centered hierarchies to be mentioned below.

(ii) Ring/Functionality. Structural features describing rings and functionality are generally in accord with chemically intuitive concepts of the significant aspects of molecular structure. Functionalities generally comprise small groups of connected atoms and bonds, while ring fragments describe ring nuclei and/or individual rings at varying levels of specification of composition and substitution. Fragments of this type have been used as part of mixed-feature sets taken from fragment codes in studies involving substructural analysis[7,25] and multiple regression, discriminant analysis studies, and cluster analysis studies, relating structure to acute toxicity, carcinogenicity, and mutagenicity.[15-17] Such variables, manually assigned, have also been used in substructural analyses by emplolying linear discriminant analysis and displaying sets of structures by principal components plots and nonlinear maps.[26] A fragment code has also provided variables for canonical correlations analyses and plots, relating structure to biological assay spectra for neuroleptics and morphinomimetics.[27] Ring/functionality variables have been derived automatically from Wiswesser line notation (WLN) in multiple-regression studies of penicillin serum binding[28] and heats of vapourization.[29] The WRAIR fragments were found to be unsuitable for one substructural analysis application because of lack of specificity of bond type. Such variables have also formed part of a "fragment code" automatically generated from WLN for structure–toxicity analyses[18] and have been generated manually from WLN records for SAR studies of odor intensity.[19]

Larger functionality descriptors derived from connection tables are incorporated in one substructural analysis system.[30] Functionalities are also included in the mixed sets of descriptors used by Jurs and co-workers,[20-24] who term this type of feature "substructures". A pattern-recognition study of antitumor activity[31] utilized structural features descriptive of ring systems, including all imbedded rings and incorporating the type and position of heteroatoms and substituent patterns in six-membered rings. Detailed ring features at several levels of specificity were used in substructural analysis studies of antitumor activity.[32,33] Analyses of WLN gave structural features representing the type and relative positions of heteroatoms and substituents in six-membered ring systems as variables in regression analyses of physicochemical thermochemical and biological properties[34-37] and in cluster analyses.[29]

Substructures of these kinds are also commonly used in group additivity schemes for molecular properties and must be derived from structural representations in the automation of such schemes, as noted above. Procedures of this sort have been described for critical constants by using fragments derived from WLN,[38] for critical properties, ideal gas functions, and heats of formation by using a connection table structure representation,[39] and for partition coefficients also based on connection tables.[5] An additive–constitutive model, with fragments at various levels of specificity, has been used for calculation of "BC(DEF)" values, from which several physicochemical properties may be calculated.[40] Ring descriptors derived from connection tables have been included in a fragmentation code for prediction of various physicochemical properties, including molar refractivity, molar volume, and parachor.[41]

(iii) Atom/Bond-Centered. Atom- and bond-centered fragments are defined in terms of the concentric area of structure surrounding each atom or bond and are delineated at an appropriate level of specificity by atom and bond types, numbers of nonhydrogen connections, etc. The use of these fragments was pioneered by Lynch's group for screen generation in substructure searching.[42,43] Examples of some of these fragments are shown in Figures 2 and 3: the way in which a natural hierarchy is formed, covering increasing extents of structural specificity, is evident. A hierarchical display of this sort, with activity contributions for each fragment, is a basis of the ABSCISSA substructural analysis system.[44]

These types of structural feature have been applied in a number of structure–activity studies. The augmented atom

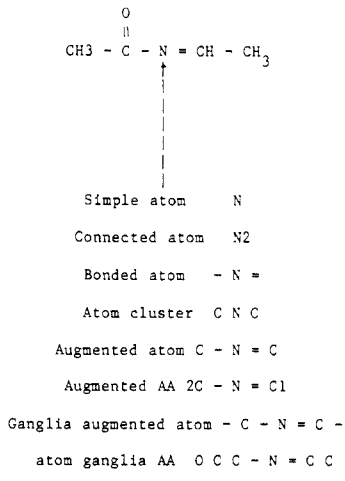

Figure 2. Atom-centred fragments.

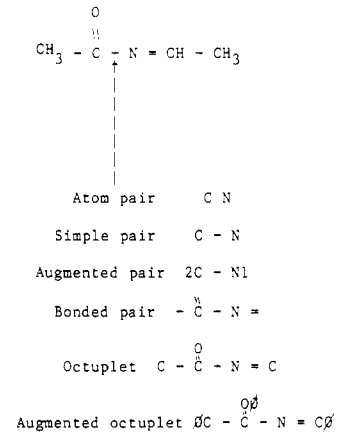

Figure 3. Bond-centred fragments.

has been most commonly used. Manual encoding of fragments of this kind have been used in pattern-recognition studies of sedatives and tranquillizers[45] and of analgesic activity.[46] Algorithmically derived, augmented atoms have been used alone as descriptors in automatic classification of chemical structures[47,48] and in multiple-regression analyses of penicillin serum binding.[49] Connected atoms, bonded atoms, and augmented atoms have been used in a comparison of cluster analysis procedures for classification of chemical structures, fragments of each type being used alone in the analyses.[50] In combination with other types of descriptors they have been applied in pattern-recognition studies of antitumor agents,[31] and in substructural analysis:[32,33,51,52] this last work also involved the next level of specificity in descriptors of this sort, the ganglia augmented atom.[33] This sort of fragment has been used in substructural analysis studies of mutagenicity.[24] Very similar features have been included in a fragment code for property prediction from additive models.[41] Atom-centered fragments of this size with varying levels of specificity of bond description have been used in pattern recognition and ISODATA clustering studies of several biological activities.[53]

The DARC/PELCO technique involves use of a hierarchy of atom-centered descriptors from the DARC structure-handling system. This technique has been applied in multiple-regression SAR analyses,[54-56] analysis of alkyl steric effects,[57] and prediction and correlation of chromatographic Kovats indices.[58] It has also been applied to determine structural variables contributing to the activity of phenylaklylamines as PNMT inhibitors[59] and hallucinogenic agents.[60]

Bond-centered fragments, by comparison, have been relatively little used in such applications. One such study contrasted the performance of simple pairs and augmented pairs as variables in the multiple-regression analysis of local anaesthetic activity of diverse compounds.[61] Augmented pairs, bonded pairs, and octuplets have each been used as variables for comparison of cluster analysis algorithms for chemical structure classification.[50] Bonded pair fragments (excluding the commonest) have also been applied in substructural analysis.[52] A system for generating fragments from the BE matrix form of connectivity record, for substructural analysis procedures, has been described.[62,63] All possible subgraphs are algorithmically derived in a hierarchical fashion and then automatically reduced in number on the basis of assessed chemical relevance and frequency of occurrence. A procedure for estimation of thermochemical parameters, based on bond-additivity assumptions with allowance for proximity effects, requires identification of constant-size bond- and atom-centered fragments of 2–4 atoms within a structure or partial structure.[4]

(iv) Paths. Other structural descriptors are based on "paths" of atoms and bonds through the structure. Heteropaths, i.e., each path joining two heteroatoms, have been used in pattern recognition analyses,[31] while "linear sequences", chains of four to six specified atoms with bonds specified only as ring or nonring, have been applied in substructural analysis.[33] The ultimate extension of this approach, the consideration of all paths in a structure, was employed for an overall assessment of structural similarity.[64]

A substructural analysis system includes descriptors comprising pairs of bonded atom fragments, with a description of their separation.[30] This "separation" is either a topological distance (the number of bonds in the shortest linking path) or a spatial distance, derived from a conformational representation.

(v) Templates. Alternative forms of descriptors are derived from what may be termed "template models". If all the compounds of a data set can be superimposed on some general structural template, variables may be generated to reflect the presence of structural units at specified positions, so preserving the context of the features, generally lost in fragmentation. Free–Wilson analysis is the most widely known example of this sort of procedure,[6] with manually encoded substituent occurrence counts on a common ring system which are used as variables in multiple-regression analysis. This technique may be extended to a rather wider range of structural types by inclusion of fusion points, heteroatoms, etc. as variables; this has been demonstrated with automatic descriptor generation.[35] Correlation equations have also been formulated for large sets of relatively diverse structures, parametrized by manually assigned structural features and/or physicochemical parameters for specific positions (see for example ref 65 and 66): these may be regarded as "loose template" models, and the application of computerized structure handling could be valuable in these circumstances.

Several papers describe template matching (usually with noncomputerized feature generation) of rather more diverse structure sets. One such used the structural units in the template matrix as input to the "perceptron" pattern-recognition technique.[67] A series of compounds with weak or strong pressor activity were similarly coded to derive variables for input to the principal components analysis in order to classify the pharmacological activities,[68] and a similar analysis was carried out with a set of diverse compounds of varied pharmacological categories:[69] in both these studies the coded positions were parametrized by molar refractivity values. In this latter case, the structures were superimposed on a "superstructure", the hypothetical parent of all the compounds in the data set: this concept is very similar to the "hyperstructure" of the DARC/PELCO system.[56]

Other template models, with positions parametrized by molar refractivity or fragment molecular connectivities (i.e.,

topological indices for partial structures), have been used for discriminant analyses and pattern-recognition studies of therapeutic categorization of steroids.[70-72] The best discrimination here was obtained by the use of "small-radius" topological indices (equivalent to parametrizations for simple or connected atoms in the atom-centered hierarchy) to denote each position in the template, rather than a more complex topological descriptor or physicochemical property. These authors also note that procedures deriving descriptors from whole-structure representations may be preferable in avoiding the "highly subjective" template-fitting process.[72]

A template approach has also been described for the assessment of the nature of essential pharmacophoric groupings.[73] This method uses simple statistical procedures, based on information theory, to evaluate relative frequencies of structural feature occurrence and was applied to fungicidal carboxamides and phenethylamine β-agonists and antagonists.

(vi) Other. The analysis of matrices reflecting distances, either topological (i.e., bond distance) or topographic (i.e., spatial distance), between structural units, in order to identify common factors in active structures has been described.[25] A simpler procedure for this purpose, the permutation of listings of linear notations in order to bring together similar substructures, has been widely used.[74]

COMPARISON OF TYPES OF STRUCTURAL FEATURES

In this section the various types of structural feature noted above are examined critically in order to illustrate the sometimes conflicting desiderata in selecting the sorts of features likely to be most useful in structure–property studies. Their usefulness must be assessed against the interlinked criteria of (i) their appropriateness as variables in the relevant statistical procedures, and (ii) the interpretability and/or predictive value of the results obtained.

(i) Feature Selection. One aspect of structural feature derivation, fundamental to the approach adopted, is the extent to which the variables used are "intuitively" derived for the particular problem at hand. At one extreme, substructures of interest would be defined after an initial examination of the data set and extracted by a substructure search; at the other, an algorithmically defined type of fragment, e.g., bonded pair or ring nucleus, could be specified. In the latter case the same type of structural feature could be used for many different studies; the actual variables so derived would, of course, vary with the composition of each set of structures.

Both these procedures are "open ended", since they may be applied to any sets of structures, and generate appropriate features; complete computer-readable structure representations are required. If the data-set structures are represented by a fragment code, the structural descriptors which may be used are "fixed", i.e., limited to those used in the code. In this case there is no flexibility either to generate open-ended algorithmically defined descriptor sets or to select appropriate data set dependent descriptors. Fixed sets of descriptors may, of course, also be generated from the total structure representations by feature perception algorithms, as in automated property estimation.[5,38,39] This approach could also be used in SAR investigations, if some standard list of substructures to be used in all cases were drawn up. It does not seem to have been adopted other than in the use of fragment codes, full structure representations being applied either to systematic algorithmic feature generation or to interactive substructure selection. However, structure–activity studies of toxicity, carcinogenicity, and mutagenicity have used a "fragment code" derived from WLN representation by the CROSSBOW set of programs.[18]

Jurs et al. argue in favor of the interactive selection of problem-dependent substructures[75,76] by the "application of common sense and the experience of the researcher".[22] On the other hand, it has been suggested that the use of algorithmically defined variables for SAR avoids possible bias due to preconceptions and may give a "new angle" to the data analysis.[61] They may be more appropriate for a "first look" at a data set and have advantages of convenience in examining large data sets.

Another point which has received some discussion is the extent to which the features used should be in accordance with "chemical intuition". It is generally agreed that it is desirable for structural variables to be "chemically meaningful" in order for the results to be readily interpretable and used predictively. Interactive selection of problem-dependent substructures by the investigator[75,76] obviously gives a high degree of chemical relevance. Algorithmically defined fragments are not, however, necessarily deficient from this point of view. Chu et al.[31] and Broome et al.[77] both chose the augmented atom fragment as a satisfactory representation of functional groups. A more complex algorithmic definition of functional group (any connected subset of atoms which does not contain carbon–carbon single or aromatic bonds) was found to be too complex and specific to be useful in one substructural analysis application.[52] The symbols of various linear notations may also be a useful basis for functionality definition.[28,78] Algorithmic feature perception routines for detecting simple or complex functionalities have also been described for computer-aided synthesis[79] and reaction mechanism studies.[9,80] These procedures are more efficient than an atom-by-atom search and could therefore be advantageous in SAR studies with very large data sets. Ring fragments, at various levels of specificity (as noted above), are both readily algorithmically derived and straightforwardly rationalized in chemical terms. The larger fragments (heteropaths, atom chains, octuplets, ganglia augmented atoms, etc.) may be valuable in defining "pharmacophoric patterns" in a single feature.[31] In general, it seems that "chemically meaningful" results can be obtained by using any type of structural feature defined as a set of connected atoms, whether algorithmically defined or preselected.

Two related aspects of variable selection are the extent to which fragments may overlap and the inclusion, or not, of all atoms and bonds in the set of features used. As a general rule, algorithmically derived variable sets tend to incorporate the whole structure, while "preselected" sets do not, though this need not necessarily be so. For property estimation it is usually essential that the whole structure be considered, so that the incremental contributions of all structured units are included. However, some physicochemical–chemical properties may be dependent on only a part of the structure, and a "fixed set" approach may be feasible. p*K*a values, for example, are predicted automatically after a feature perception generation of functional groups on a fixed list.[9] For correlation with biological activity, classification, and pattern searching the total structure need not necessarily be considered since here the aim is to identify those parts of the structure contributing to activity.

The question of overlap of structures arises in several ways. If several types of descriptor, e.g., rings and augmented atoms, or more than one level of atom- or bond-centered fragments are combined in a variable set, then particular problems of redundancy arise: this is discussed later. Some overlap is always present: if H_2N–CH_2–COOH is fragmented into the "discrete" NH_2, CH_2, and COOH groups there is, in a sense, an overlap of the N–C and C–C bonds. The problem is particularly noticeable with larger atom- and bond-centered fragments, where each atom or bond is the center of a frag-

ment and will also appear in several other fragments, so that each aspect of the structure is represented in several ways. This may be undesirable statistically, because of the increase in number of variables and of intervariable correlation. It may also cause difficulty in interpretation: a comparison of analyses of serum binding of penicillins with overlapping[49] and nonoverlapping[28] fragments has been made from this viewpoint. Conversely, the greater variety of structural descriptors provided by overlapping may be valuable in pattern-searching substructural analysis, where statistical criteria may not be so rigorous.

(ii) Statistical Aspects. Two particular statistical problems impinge upon the choice of structural descriptors. These are the ratio of observations (compounds) to variables (descriptors)—in pattern recognition terms the "sample/feature ratio"—and relationships between the descriptors, which will deviate from statistical independence to some greater or lesser extent, leading to redundancy.

A considerable amount has been written on the minimum desirable ratio of observations to variables in statistical techniques involving some multivariate function, e.g., multiple regression, discriminant analysis, and linear learning machines. The problem does not arise in the same way with techniques of classification (including principal components and factor analysis), display, and pattern searching. This is still an area of some controversy, which cannot be fully dealt with here, but as a general guide we may note that for some types of pattern recognition the ratio should be at least 3:1 for learning machines,[81] while Topliss and Costello recommend rather higher ratios for multiple regression.[71] Certainly this problem becomes evident if the number of variables approaches the number of observations, as may happen, in particular, with larger algorithmically derived substructures. This does not imply that the results are worthless in such cases but rather that they must be regarded with considerable caution, in view of possible spurious values. In the case of multiple-regression analysis, for example, a low observation/variable ratio would imply that the overall statistical validity of the equation produced would be in doubt and that individual coefficient values should be regarded critically: such an analysis would be valid as a exploratory data-analytic procedure, examining overall trends, rather than an exact hypothesis testing technique.[35] Ways of minimizing this problem have included the use of only one type of atom- or bond-centered descriptor in an analysis,[33,50,61] the use of stepwise analytical procedures to select the most significant subset of variables[15,19,28,34,35,49,61] (although this approach does not entirely avoid the problem[82], and techniques for preprocessing and feature selection, which will be discussed below. Monte Carlo studies, assessing the likelihood of spurious correlation by testing against random dependent variables, may also be appropriate.[83]

The problem of statistical association and redundancy among descriptor variables, which may affect both the statistical and interpretative aspects of an analysis, was first noted in this connection by Adamson et al.[84] This problem may be dealt with to some extent in some of the same ways, e.g., restricting the types of descriptor, using a stepwise analysis to omit highly correlated variables, or adopting a preprocessing technique. Hodes describes the reduction of redundancy due to overlap among ganglia augmented atom descriptors by removing the smaller substructures of this type[33] and suggests more complicated techniques for removal of such redundancy.[52]

Preprocessing techniques aim either to select a subset of the total set of descriptors on statistical grounds or to create new variables by a transformation process. The purpose may be to reduce redundancy among the variables or to improve the performance of a descriptor set for a given problem. Jurs et al. suggest that the interactive selection of substructures by the investigator will be the most effective preprocessing technique and discount the practicability of statistically based feature selection.[75,76] Some procedures of this kind have, however, been reported. Chu et al. utilize the weight–sign change feature selection technique in a pattern-recognition study,[31] while Wijnne describes the use of principal components analysis to transform a set of augmented atom descriptors.[46,85] This method gives a set of nonredundant variables, which represent aspects of molecular structure, and may be related directly to the original variables, aiding interpretability. Since the first few components will generally reflect a high proportion of the variance, this technique may also be used to reduce the total number of variables, and its wider application to studies of this sort could be worthwhile.

The appropriate size or complexity of the substructures used must also be considered, especially with algorithmically defined features. This aspect is highly interrelated with the statistical procedure being used. Equivalent results may be obtained either with a complex function of relatively simple defined fragments or with a simpler function of larger and hence more complex structural features. For example, Adamson and Bush show that a quadratic regression function (which allows, in an approximate fashion, for fragment interaction) using simple pairs gives very similar results to a linear regression with the larger augmented pair descriptor for correlation of structure with local anaesthetic activity.[61] In such cases, the use of more complex descriptors with a simpler multivariate function is usually to be preferred, for ease of interpretation and reliability of extrapolation.

A general problem arises with regard to interpretability and generalizability of results if relatively simple structural descriptors are used. Although such features have advantages of ease of generation and statistical amenability, their crudeness of defintion can lead to misinterpretation of the significance of the results. An early example of pattern recognition in SAR[13] was criticized on the grounds that the simple counts and ratios of S occurrence, which were significant variables, disguised the precise nature of the active substructure, a 6-S purine:[14] the results were not in error per se but were expressed in a form making reliable extrapolation difficult. For this reason it may be generally desirable to use structural descriptors as explicit, i.e., complex, as is feasible on statistical grounds.

With regard to prediction of activity, use of complex features will generally give a better correlation and hence the possibility of better prediction. However, the large number of variables involved presents two problems. First, the coefficients associated with particular substructures may be based on few observations and hence may be unreliable in prediction. Second, there may well be no value available for substructures present in the compound to be predicted. For these reasons it is usually necessary to use a less complex type of descriptor for prediction. In automated Free–Wilson analysis, for example, where descriptors vary from substituent occurrence to relative positions and to systematic inclusion of interaction terms, the less complex variables are more effective for prediction.[35,36]

In dealing with this conflict between simple and complex descriptor types, the hierarchical atom- and bond-centered fragments may have advantages. Consideration of the hierarchies present could allow a more accurate assessment of the extent of the substructures responsible for activity than with a single descriptor type and hence more reliable extrapolation of results.[44] For prediction, if values for the more complex substructures are not available, "dropping down" the hierarchy could enable a reasonable approximation: this process has been used in thermochemical property estimation.[4] In calculation of "BC(DEF)" molecular parameters, fragments

at a high (detailed) level of a hierarchy are used for "constitutive" corrections of the basic additive model.[40]

The frequency distribution of substructural descriptors over the data set under investigation is also of importance. On purely statistical grounds, it would be desirable for the variables to be roughly equifrequent for correlation. Highly skewed distributions present a number of problems: for example, slight variations in the occurrence of descriptors across activity groups tend to appear unwarrantedly significant if the descriptors are of high occurrence, while, conversely, a long "tail" of infrequently occurring fragments presents its own problems. Singly occurring features cause trouble in regression studies, since they effectively "fix" the estimated value of their compounds, and they were omitted from a pattern-recognition study.[31] Fragment codes, being designed basically for information retrieval, would not usually include fragments of very high or very low incidence: procedures for producing sets of roughly equifrequent atom- or bond-centered fragments from connection tables are available.[43] In general, fragments of extremes of incidence would probably not be used in interactively generated sets.

However, there is, in all procedures for removing high- and low-incidence features, a danger that valuable information may be lost. Distributions of particular types of fragments are determined by the composition of the data set. High-incidence fragments, although they may very well be common "background noise" features, may also represent some important factors in the data-set composition which should not be overlooked. Low-incidence features associated with activity may be of particular importance in pointing to novel structural classes: one major role for automatic structural analysis is to pick these out from large data sets, and this opportunity may be missed if low-incidence fragments are ignored. One substructural analysis methodology emphasizes activity contributions from infrequent features for this reason.[33]

For classification and display techniques, standardization of variables may be used to counteract skew distributions. However, since the scales of all the variables are the same (i.e., occurrence counts), the effect will be a weighting in favor of low-incidence variables, affecting the overall classification, which may or may not be desired. Cluster analysis by structural variables shows this point clearly.[29]

In general, any technique aimed at altering the frequency distribution of the variables must be used with caution and with a particular view to the desired effect of low-incidence fragments in the analysis.

(iii) Descriptor Combination. Sets of structural descriptors are commonly used alone in structure–property studies. Sometimes, however, other types of variables have been added, giving mixed sets of descriptors. There are two types of variable which have been used in this way.

First, there are topological indices, single numbers or sets of numbers which represent aspects of molecular structure and which are derived from structure diagram representations by formalisms and parametrizations based on graph theory, information theory, or simple pragmatism (see, for example, ref 86–90). Such indices have been applied in various areas of SAR studies, but a detailed consideration of their properties is not possible here. Some indices, applied to both total and partial molecular structures, have been used in pattern-recognition studies, in conjunction with other types of descriptor.[20-23,75,76]

Second, physicochemical properties, either measured or estimated, may be used in conjunction with substructural descriptors, in an analogy to the mixed Hansch–Free–Wilson technique of "classical" QSAR.[91] The partition coefficient has been used with substructural descriptors in regression, discriminant, and cluster analyses of acute toxicity[15-17] and in pattern-recognition studies of carcinogenicity of polycyclic hydrocarbons while electronic indices from simple molecular orbital calculations have also been used in combination.[23] Parameters representing steric bulk, or molecular shape, may be similarly applied.

The mixed-descriptor approach may have advantages in analyzing complex multifaceted data sets, but it presents a number of problems. In particular, interrelationships between variables become particularly complicated, since the substructural fragments contribute to the physicochemical or topological parameters: it has been noted that different types of descriptor may "substitute" for one another with some data sets.[21] Problems in interpretability of results with mixed descriptor sets may become extreme[20,21] and prediction of novel active structures on their basis difficult.[20] Dealing with problems of this sort, so as to make best use of the information provided by different types of descriptor, could be a fruitful area for future work. In particular, the relative merits, appropriate application, and interrelationships of the use of counts of structural fragments and of topological indices and other structural parametrizations need further elaboration. Both sorts of descriptor are readily derivable automatically from computer-readable structure representations, and both encode, in their own way, structural information. In general, molecular indices reduce the number of potential variables and hence alleviate some statistical problems, but interpretation of the results is frequently difficult, particularly if the parametrization of the indices has been largely empirical. Further theoretical studies and comparative evaluations in this area would be valuable.

IMPLICATIONS FOR CHEMICAL INFORMATION SYSTEMS

The computer-based systems which have supported the studies described above have varied from large operational systems deisgned for storage and retrieval of organizations' data[7,30,32,77] to smaller scale systems specifically devised for computer-aided SAR studies.[35,60,76] Nonetheless, some general conclusions may be drawn.

First, a total structure representation, though not essential, is highly desirable for work in this area. Fixed fragment codes, though they may give useful results,[7] are too limiting in the sort of structural features which may be identified and hence in the descriptors which may be used. The problems may be alleviated, to some extent, by careful choice of fragments to suit the kinds of compounds in the file; but fragment coding denies the flexibility of descriptor generation usually required in SAR studies.

Given the desirability of total structure representations the choice between linear notations and full connectivity representations is essentially one of convenience. In these applications, this is largely dependent on the type of compounds being studied and the type of descriptors to be derived. Notations have evident advantages in selecting "chemically significant" descriptors (rings, substituent interactions, functional groups, etc.), since these moieties are clearly and directly encoded in notations such as WLN.[28,34,35] Notations would also be paricularly convenient for dealing with similar structures, e.g., derivatives of a common parent. However, connectivity tables have considerable advantages of flexibility and generality over notations, and it will usually be easier to derive rings and functionalities from connection tables[79] than to implement generalized descriptor identification from linear notations. A system for computer-aided SAR should therefore either be connection table based or have good notation-to-connectivity interconversion facilities.

It is worth noting that the amount of structural detail explicitly present in the structure representations will often need

to be greater for SAR application than for storage and retrieval. Such factors as distinction between ring and chain bonds in connectivity records and accurate specification of bond types in aromatic or tautomeric systems may require more careful consideration if SAR application is envisaged than would otherwise be the case. In general, the more detail which can be included in the initial structure representation the more effectively and efficiently can descriptor generation be carried out. There may also be a need for a number of representations, carrying complementary information explicitly encoded, with ready interconversion.

The overall influence of SAR analysis in chemical information systems is likely to be in trends toward more "user friendliness" in input and output and toward greater integration, both of chemical structures and biological and other properties within data sets and of structure-handling procedures with statistical and other analysis programs. Since the sizes of structure files to be handled for SAR will be considerably less than those frequently dealt with for storage and retrieval (with the exception of large-scale substructural analyses), criteria of efficiency and effectiveness are likely to be substantially different for the two applications.

CONCLUSIONS

The discussion above has indicated the increasing use of computerized techniques of structure handling in several types of SAR applications. With the increasing availability of computing power for chemical and biochemical research and the growing tendency toward computer storage of structure–property data, the use of such techniques seems likely to become more widespread. This may happen in two ways in particular. First, computer-aided methods could be incorporated, on a routine basis, into smaller scale structure–activity and structure–property studies. Second, large scale operations (such as substructural analysis studies of biological screening data, automatic prediction of physicochemical or thermochemical properties, or structural classifications for information analysis and retrieval within large files) will gain wider use.

Such wider use will necessitate a better understanding of some aspects of these techniques, requiring both theoretical studies and comparative evaluations. In particular, descriptor selection, relationships between types of descriptor and statistical techniques, and optimal combinations of types of descriptor are aspects requiring further investigation. Additionally, new feature-perception procedures to deal with molecular properties such as hydrogen bonding could be valuable. As an indication of the possibilities, an algorithmic approach to the identification of aromaticity and tautomerism has been reported[92] as part of a system for the prediction of reaction mechanisms and products. Studies in this area could lead to a better understanding of where these relatively simple techniques, based essentially on the chemical structure diagram, must give way to explicit encoding or calculation of electronic, steric, or physicochemical parameters.

The main effects of extending the use of these techniques seem likely to be in moves toward systems using more detailed full-structure representations or perhaps a range of complementary representations with ready interconversion and toward more user-friendly systems.

A further important effect is likely to be on the quality of data included in computerized systems. Routine use of the data-analytic techniques is likely to show up inadequacies and inconsistencies in data encoding, much more than would simple storage and retrieval. One highly desirable consequence of the more general use of these methods may be a greater emphasis on data validation and more careful consideration of the nature of the data stored.[93] This in turn could help to transform data banks from a static archival function to a more dynamic role as a research tool.

REFERENCES AND NOTES

(1) Lynch, M. F.; et al. "Computer Handling of Chemical Structure Information"; Macdonald/Elsevier: London, 1971.
(2) J. E.; Ash, Hyde, E., Eds. "Chemical Information Systems"; Ellis Horwood: Chichester, 1975.
(3) Janz, G. "Thermodynamic Properties of Organic Compounds"; Academic Press: 1967.
(4) Gasteiger, J. "Automatic Estimation of Heats of Atomization and Heats of Reaction". *Tetrahedron* **1979**, *35*, 1419–1426.
(5) Chou, J. T.; Jurs, P. C. "Computer-Assisted Computation of Partition Coefficients from Molecular Structures Using Fragment Constants". *J. Chem. Inf. Comput. Sci* **1979**, *19*, 172–178.
(6) Kubinyi, H.; Kehrhahn, O. "Quantitative Structure–Activity Relationships. 3. A Comparison of Different Free-Wilson Models". *J. Med. Chem.* **1976**, *19*, 1040–1049.
(7) Cramer, R. D.; Redl, G.; Berkoff, C. E. "Substructural Analysis. A Novel Approach to the Problem of Drug Design". *J. Med. Chem.* **1974**, *17*, 533–535.
(8) Roos-Kozel B. L.; Jorgensen, W. L. "Computer-Assisted Mechanistic Evaluation of Organic Reactions. 2. Perception of Rings, Aromaticity and Tautomers". *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 101–111.
(9) Salatin, T. D.; Jorgensen, W. L. "Computer-Assisted Mechanistic Evaluation of Organic Reactions. 1. Overview". *J. Org. Chem.* **1980**, *45*, 2043–2051.
(10) Gray, N. A. B.; et al. "Stereochemical Substructure Codes for ^{13}C Spectral Analysis". *Org. Magn. Reson.* **1981**, *15*, 375–389 and references therein.
(11) Haraki, K. S. Venkataraghavan, R.; Mclafferty, F. W. "Predictions of Substructures from Unknown Mass Spectra by the Self-Training Interpretive and Retrieval System". *Anal. Chem.* **1981**, *53*, 386–392.
(12) Dubois, J.-E.; Doucet, J.-P. "Alkyl Substituent Shifts in 13C NMR Spectra of Alkynes and Alkynols. Part 1". *J. Chem. Res., Synop.* **1980**, 82–83.
(13) Kowalski, B. R.; Bender, C. F. "The Application of Pattern Recognition to Screening Prospective Anticancer Drugs. Adenocarcinoma 755 Biological Activity Test". *J. Am. Chem. Soc.* **1974**, *96*, 916–918.
(14) Unger, S. H. "Discussion of Pattern Recognition". *Cancer Chemother. Rep., Part 2* **1974**, *4*, 45–46.
(15) Enslein, K.; Craig, P. N. "A Toxicity Estimation Model". *J. Environ. Pathol. Toxicol.* **1978**, *2*, 115–121.
(16) Enslein, K. In "Structural Correlates of Carcinogenesis and Mutagenesis"; I. M., Asher, C., Zervos, Eds.; Office of Science, Food and Drug Administration: Washington, DC, 1977.
(17) Craig, P. N.; Waite, J. H. "Analysis and Trial Applications of Correlation Methodologies for Predicting Toxicity of Organic Chemicals". Report No. EPA-560/1-76-006; Environmental Protection Agency: Washington, DC, 1976.
(18) Enslein, K.; Craig, P. N. "Status report on Development of Predictive Models of Toxicological Endpoints"; Genesee Corp.: Rochester, NY, 1979.
(19) Dravnieks, A. "Correlation of Odor Intensities and Vapour Pressures with Structural Properties of Odorants". *ACS Symp. Ser.* **1977**, *No. 51*, 11–28.
(20) Stuper, A. J.; Jurs, P. C. "Structure–Activity Studies of Barbiturates Using Pattern Recognition Techniques". *J. Pharm. Sci.* **1978**, *67*, 745–751.
(21) Brugger, W. E.; Jurs, P. C. "Extraction of Important Molecular Features of Musk Compounds Using Pattern Recognition Techniques". *J. Agric. Food Chem.* **1977**, *25*, 1158–1164.
(22) Jurs, P. C.; Chou, J. T.; Yuan, M. "Computer-Assisted Structure–Activity Studies of Chemical Carcinogens. A Heterogeneous Data Set". *J. Med. Chem.* **1979**, *22*, 476–483.
(23) Chou, J. T.; Jurs, P. C. "Computer-Assisted Structure–Activity Studies of Chemical Carcinogens. An N-Nitroso Compound Data Set". *J. Med. Chem.* **1979**, *22*, 792–797.
(24) Yuta, K.; Jurs, P. C. "Computer-Assisted Structure–Activity Studies of Chemical Carcinogens. Aromatic Amines". *J. Med. Chem.* **1981**, *24*, 241–251.
(25) Golender, V. E.; Rozenblit, A. B. In "Drug Design"; Ariens, E. J., Ed.; Academic Press: New York, 1980.
(26) de Winter, M. L. "Significant Fragment Mapping, Lead Generation by Substructural Analysis". Proceedings of the 4th European Symposium on QSAR, Bath, England, 1982.
(27) Lewi, P. J. In "Drug Design"; Ariens, E. J., Ed.; Academic Press: 1980; Vol. 10, pp 308–342.
(28) Adamson, G. W.; Bawden, D. "A Method of Structure–Activity Correlation using Wiswesser Line Notation". *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 215–220.
(29) Adamson, G. W.; Bawden, D. "Comparison of Hierarchical Cluster Analysis Techniques for Automatic Classification of Chemical Structures". *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 204–209.
(30) Venkataraghaven, R. "An Integrated Data-Base System [Lederle laboratories]". PMA Science Information Subsection Annual Meeting, 1980.
(31) Chu, K. C.; et al. "Pattern Recognition and Structure–Activity Relationship Studies. Computer-Assisted Prediction of Antitumour Activity in Structurally Diverse Drugs in an Experimental Mouse Brain Tumour

System". *J. Med. Chem.* **1975**, *18*, 539–545.
(32) Hodes, L.; et al. "A Statistical-Heuristic Method of Automated Selection of Drugs for Screening". *J. Med. Chem.* **1977**, *20*, 469–475.
(33) Hodes, L. "Computer-Aided Selection of Novel Antitumor Drugs for Animal Screening". *ACS Symp. Ser.* **1979**, *No. 112*, 583–602.
(34) Adamson, G. W.; Bawden, D. "An Empirical Method of Structure–Activity Correlation for Polysubstituted Cyclic Compounds Using Wiswesser Line Notation". *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 161–165.
(35) Adamson, G. W.; Bawden, D. "A Substructural Analysis Method for Structure–Activity Correlation of Heterocyclic Compounds Using Wiswesser Line Notation". *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 164–171.
(36) Adamson, G. W.; Bawden, D. "Automated Additive Modeling Techniques Applied to Thermochemical Property Estimation". *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 242–246.
(37) Adamson, G. W.; Bawden, D. "Substructural Analysis Techniques for Empirical Structure–Property Correlation. Application to Stereochemically Related Molecular Properties". *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 97–100.
(38) Brasie, W. C.; Liou, D. W. "Chemical Structure Coding". *Chem. Eng. Prog.* **1965**, *61* (5), 102–108.
(39) Jochelson, N.; Mohr, C. M.; Reid, R. C. "The Automation of Structural Group Contribution Methods in the Estimation of Physical Properties". *J. Chem. Doc.* **1968**, *8*, 113–122.
(40) Cramer, R. D. "BC(DEF) Parameters. 2. An Empirical Structure-Based Scheme for the Prediction of Some Physical Properties". *J. Am. Chem. Soc.* **1980**, *102*, 1849–1859.
(41) Osinga, M.; Verrijn Stuart, A. A. "Documentation of Chemical Reactions. IV. Further Applications of WLN Analysis Programs: A System for Automatic Generation and Retrieval of Information on Chemical Compounds (AGRICC)". *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 26–32.
(42) Crowe, J. E.; Lynch, M. F.; Town, W. G. "Analysis of Structural Characteristics of Chemical Compounds in A Large Computer-Based File. Part 1. Noncyclic Fragments". *J. Chem. Soc. C* **1970**, 990–996.
(43) Adamson, G. W.; et al. "Strategic Considerations in the Design of a Screening System for Substructure Searches of Chemical Structure Files". *J. Chem. Doc.* **1973**, 13, 153–157 and references therein.
(44) Bawden, D., paper in preparation.
(45) Chu, K. C. "Application of Artificial Intelligence to Chemistry. Use of Pattern Recognition and Cluster Analysis to Determine the Pharmacological Activity of Some Organic Compounds". *Anal. Chem.* **1974**, *46*, 1181–1187.
(46) Wijnne, H. In "Quantitative Structure–Activity Analysis"; Franke, R., Oehme, P., Eds.; Academic-Verlag: West Berlin, 1978.
(47) Adamson, G. W.; Bush, J. A. "A Method for the Automatic Classification of Chemical Structures". *Inf. Storage Retr.* **1973**, *9*, 561–568.
(48) Adamson, G. W.; Bush, J. A. "A Comparison of the Performance of Some Similarity and Dissimilarity Mesaures in the Automatic Classification of Chemical Structures". *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 55–58.
(49) Adamson, G. W.; Bush, J. A. "Method for Relating the Structure and Properties of Chemical Compounds". *Nature (London)* **1974**; *248*, 406–407.
(50) Willett, P. "A Comparison of Some Hierarchical Agglomerative Clustering Algorithms for Structure–Property Correlation". *Anal. Chim. Acta* **1982**; *136*, 29–37.
(51) Hodes, L. "Computer-Aided Selection of Compounds for Antitumour Screening: Validation of a Statistical Heuristic Method". *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 128–132.
(52) Hodes, L. "Selection of Molecular Fragment Features for Structure–Activity Studies in Antitumour Screening". *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 132–136.
(53) Sacco, W.; et al. "Introduction to Pattern Recognition Applications in Structural Correlates of Carcinogenesis and Mutagenesis"; Asher, I. M., Zervos, C., Eds.; Office of Science, Food and Drug Administration: Washington, DC, 1977, and references therein.
(54) Dubois, J. E.; Laurent, D.; Bost, P.; Chamband, S.; Mercer, C. "DARC System. DARC/PELCO Method. Strategies of Correlation Research Applied to a Group of Antiinfluenza Adamantanamines (in French)". *Eur. J. Med. Chem.* **1976**, *11*, 225–236.
(55) Duperray, B.; Chastrette, M.; Cohen-Mahabeh, M.; Pachew, I. H. "Analysis of Bactericidal Activity of Aliphatic Alcohols and β-Naphthols by Hansch and DARC-PELCO Methods (in French)". *Eur. J. Med. Chem.* **1976**, *11*, 433–437.
(56) Mercer, C.; Dubois, J. E. "Comparison of Molecular Connectivity and DARC/PELCO Methods: Performance in Antimicrobial Halogenated Phenol QSARs". *Eur. J. Med. Chem.* **1979**, *14*, 415–423.
(57) Panage, A.; Macphee, J. A.; Dubois, J. E. "Relationship between Topology and the Steric Parameter Es". *Tetrahedron* **1980**, *36*, 759–768.
(58) Chretien, J. R.; Dubois, J. E. "New Perspectives in the Prediction of Kovats Indices". *J. Chromatogr.* **1976**, *126*, 171–189.
(59) Mercier, C.; Sobel, Y.; Dubois, J.-E. "Methode DARC/PELCO: QSAR Unique de Phenyl Alkylamnes Inhibitrices de la PNMT". *Eur. J. Med. Chem.* **1981**, *16*, 473–476.
(60) Sobel, Y.; Mercier, C.; Dubois, J.-E. "Methode DARC/PELCO: QSAR de Phenylalkylamines Methoxylees Hallucinogenes". *Eur. J. Med. Chem.* **1981**, *16*, 477–479.
(61) Adamson, G. W.; Bush, J. A. "Evaluation of an Empirical Structure–Activity Relationship for Property Prediction in a Structurally Diverse Group of Local Anaesthetics". *J. Chem. Soc., Perkin Trans. 1* **1976**, 168–172.
(62) Friedrich, J.; Ugi, I. "Substructure Searching and Structure Property Locating by Means of Subgraph Generation". *MATCH* **1979**, *6*, 201–211.
(63) Friedrich, J.; Ugi, I. "Substructure Retrieval and the Analysis of Structure–Activity Relations on the Basis of a Complete and Ordered Set of Fragments". *J. Chem. Res., Synop.* **1980**, 70.
(64) Randic, M.; Brissey, G. M.; Spencer, R. B.; Wilkins, C. L. Use of Self-Avoiding Paths for Characterization of Molecular Graphs with Multiple Bonds". *Comput. Chem.* **1980**, *4*, 27–43.
(65) Hansch, C.; Silipo, C.; Steller, E. E. "Formulation of de novo Substituent Constants in Correlation Analysis: Inhibition of Dihydrofolate Reductase by 2,4-Diamino-5-(3,4-dichlorophenyl)-6-substituted Pyrimidines". *J. Pharm. Sci.* **1975**, *64*, 1186–1191.
(66) Kim, K. H.; Hansch, C.; Fukunga, J. Y.; Steller, E. E.; Jow, P. T. C.; Craig, P. N.; Page, J. "Quantitative Structure–Activity Relationships in 1-Aryl-2-(alkylamino)ethanol Antimalarials". *J. Med. Chem.* **1979**, *22*, 366–391.
(67) Hiller, S. A.; Golender, V. E.; Rosenblit, A. B.; Rastrigin, L. A.; Glaz, A. B. "Cybernetic Methods of Drug Design. 1. Statement of the Problem–The Perceptron Approach". *Comput. Biomed. Res.* **1973**, *6*, 411–421.
(68) Cammarata, A.; Menon, G. K. "Pattern Recognition Classification of Therapeutic Agents According to Pharmacophore". *J. Med. Chem.* **1976**, *19*, 739–748.
(69) Menon, G. K.; Cammarata, A. "Pattern Recognition. II. Investigation of Structure–Activity Relationships". *J. Pharm. Sci.* **1977**, *66*, 304–314.
(70) Henry, D. R.; Block, J. H. "Classification of Drugs by Discriminant Analysis using Fragment Molecular connectivity". *J. Med. Chem.* **1979**, *22*, 465–472.
(71) Henry, D. R.; Block, J. H. "Pattern Recognition of Steroids using Fragment Molecular Connectivity". *J. Pharm. Sci.* **1980**, *69*, 1030–1034.
(72) Henry, D. R.; Block, J. H. "Steroid Classification by Discriminant Analysis Using Fragment Molecular Connectivity". *Eur. J. Med. Chem.* **1980**, *15*, 133–138.
(73) Hubel, S.; Rosner, T.; Franke, R. "The Evaluation of Topological Pharmacophores by Heuristic Approach." *Pharmazie* **1980**, *35*, 424–433.
(74) Saggers, D. T. "The Application of the Computer to a Pesticide Screening Programme". *Pestic Sci.* **1974**, *5*, 341–352.
(75) Brugger, W. E.; Stuper, A. J.; Jurs, P. C. "Generation of Descriptors from Molecular Structure". *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 105–110.
(76) Stuper, A. J.; Brugger, W. E.; Jurs, P. C. "A Computer System for Structure–Activity Studies using Chemical Structure Information Handling and Pattern Recognition Techniques". *ACS Symp. Ser.* **1977**, *No.* 52.
(77) Broome, P. H.; et al. "Pattern Recognition Applications in Chemistry and Pharmacology. IV. Generation of Atom-Centered Fragments of Organic Compounds by Computer". Report AD-A051 936; NTIS: 1977.
(78) Lin, C. H. "Chemical Inference based on SEFLIN. 1. Basic Cognizance of Molecular Shape, Fragments, and Atomic Environment of Organic Compounds". *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 47–51.
(79) Bersohn, M.; Esack, A. "Functional Group Discovery Using the Concept of Central Atoms". *Chem. Scr.* **1976**, *9*, 211–215.
(80) Wipke, W. T.; Howe, W. J. "Computer-Assisted Organic Synthesis". *ACS Symp. Ser.* **1977**, *No. 61*.
(81) Soltzberg, L. J.; Wilkins, C. L.; Kaberline, S. L.; Lam, T. F.; Brunner, T. L. "Evaluation and Comparison of Pattern Classifiers for Chemical Applications: Adaptive Digital Learning Networks and Linear Discriminants". *J. Am. Chem. Soc.* **1976**, *98*, 7144–7151.
(82) Topliss, J. G.; Edwards, R. P. "Chance Factors in Studies of Quantitative Structure–Activity Relationships". *J. Med. Chem.* **1979**, *22*, 1238–1244.
(83) Kier, L. B.; Hall, L. H. "Structure–Activity Studies on Hallucinogenic Amphetamines Using Molecular Connectivity". *J. Med. Chem.* **1977**, *20*, 1631–1636.
(84) Adamson, G. W.; Lambourne, D. R.; Lynch, M. F. "Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-Based File. Part III. Statistical Association of Fragment Incidence". *J. Chem. Soc., Perkin Trans. 2* 1972, 2428–2433.
(85) Wijnne, H. In "Biological Activity and Chemical Structure"; Keverling Buisman, J. A., Ed.; Elsevier: Amsterdam, 1977.
(86) Rouvray, D. H. "The Search for Useful Topological Indices in Chemistry". *Am. Sci.* **1973**, *61*, 729–735.
(87) Kier, L. B.; Hall, L. C. "Molecular Connectivity in Chemistry and Drug Research"; Academic Press: New York, 1976.
(88) Bonchev, D.; Trinajstic, N. "Information Theory, Distance Matrix, and Molecular Branching". *J. Chem. Phys.* **1977**, *67*, 4517–4533.
(89) Mekenyan, O.; Bonchev, D.; Trinajstre, N. "Chemical Graph Theory: Modeling the Thermodynamic Properties of Molecules". *Int. J. Quantrum Chem.* **1980**, *18*, 369–380.

(90) Sabljic, A.; Trinajstic, N. "Quantitative Structure–Activity Relationships: The Role of Topological Indices". *Acta Pharm. Jugosl.* **1981**, *31*, 189–214.
(91) Kubinyi, H. "Quantitative Structure–Activity Relationships. 2. A Mixed Approach, Based on Hansch and Free-Wilson Analysis". *J. Med. Chem.* **1976**, *19*, 587–600.
(92) Rispin, A. "Introduction to Symposium on the Development and Use of Reliable Data Bases for Quantitative Structure–Activity Relationships". *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 1, and references therein.
(93) Tinker, J. "Relating Mutagenicity to Chemical Structure". *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 3–7.

Some Heuristics for Nearest-Neighbor Searching in Chemical Structure Files

PETER WILLETT

Department of Information Studies, University of Sheffield, Western Bank, Sheffield S10 2TN, England

Received August 17, 1982

Nearest-neighbor searching usually involves inspecting all of the records in a file for the record which best matches an input query. Three heuristics are described for nearest-neighbor searching in chemical structure files where molecules are represented by fragment bit strings. The procedures can reduce the number of compounds inspected by between 81% and 97%, depending upon the heuristic and the matching function used, while still ensuring the identification of the nearest neighbor.

INTRODUCTION

Exact match and partial, or inclusive, match searching algorithms are widely used in computer-based chemical information systems for the purpose of registration and substructure search, respectively. A less common facility is provision for best-match, or nearest-neighbor, searches in which the structure or structures most similar to an input query structure are retrieved, similarity being defined on the basis of some similarity coefficient or distance function[1] which reflects the number of fragments common to the query and to a molecule in the file. Best-match searching forms the basis for the k nearest-neighbor classification[2] and plays an important role in the use of spanning trees[3,4] and of automatic classification techniques.[5-9]

The general problem of finding best matches is defined by Friedman et al.[10] as "...given a file of N records (each of which is described by k real valued attributes) and a dissimilarity measure D, find the m records closest to a query record (possibly not in the file) with specified attribute values". The obvious, brute-force algorithm for best-match searching is to compute the distance between the query and each of the records in the file and then to select the m shortest distances; this algorithm has a file size dependency of $O(N)$ and is much too time-consuming for all but the smallest files.

This paper describes the use of several heuristics which, although not reducing the complexity of the search below $O(N)$, are sufficiently powerful to allow nearest-neighbor searches of chemical structure files to be carried out at reasonable computational cost. All of the experiments consider the retrieval only of the nearest neighbor, i.e., $m = 1$, but the procedures outlined may be generalized to a range of other closest point problems for which $m > 1$.

NEAREST-NEIGHBOR SEARCHING

An efficient nearest-neighbor algorithm will be one which avoids the calculation of most of the distances while still calculating the distances for those few records which are, in fact, near the query structure. Several types of criteria have been suggested to reduce the number of calculations required, including the projection of the d-dimensional records onto a lower dimensional space where most of the distance calculations are performed[11,12] and grouping records into clusters so that several records may be searched, or eliminated from the search, simultaneously.[10,13-16] Many of the cited algorithms may not be directly applicable to best-match searching in a chemical context since they assume that the attributes are continuous variables, whereas chemical structures are usually characterized by a binary fragment description. In this, each of the structures in a file is represented by a bit string in which the i'th bit is set if the corresponding fragment is present in the structure. Also, it is often assumed that the records lie in a d-dimensional space where d is small, typically 2 or 3, so that multiplicative terms in d in the equation describing the number of matches required may be neglected; in a chemical structure system, d may be of the order of 10^2 or 10^3 (the number of bits in the bit string), and such algorithms are, accordingly, quite impracticable. Thus the $O(\log N)$ procedure due to Friedmam et al.[10] involves a constant of proportionality of about 1.6^d while the search method of Bentley et al.[16] involves the inspection of all of the $3^d - 1$ cells adjacent to a given cell in a d-dimensional space. Marimont and Shapiro[17] discuss the dimensionality problem, but their experiments were still restricted to spaces with $d \leq 40$.

van Marlen and van den Hende[18] and Rasmussen et al.[19] have described best-match retrieval algorithms for use with machine-readable mass spectra files, where a structure is characterized by a bit string corresponding to the peaks observed in the molecular mass spectrum, while Smeaton and van Rijsbergen,[20] Murtagh,[21] and Perry[22] have studied best-match searching in the context of document retrieval systems. Smeaton and van Rijsbergen note that an inverted file may be used to increase search efficiency since a query needs to be matched only against those documents with which it has at least one term in common. They then describe experiments with an upper-bound procedure which enables a best match search to be teminated before all of the documents in the inverted file lists corresponding to a query have been inspected. Murtagh and Perry describe an extension of this algorithm in which additional upper bounds are calculated, this resulting in a further reduction in the number of documents that need to be matched against a query.

EXPERIMENTAL DETAILS AND SIMILARITY MEASURES

The experiments used a set of 2335 structurally disparate compounds from the Index Chemicus Registry System. Each

0095-2338/83/1623-0022$01.50/0