Rapid Construction of Data Tables for Quantitative Structure-Activity Relationship Studies

John S. Delaney, Anne Mullaley, Graham W. Mullier, Graham J. Sexton, Robin Taylor,* and Russell C. Viner

ICI Agrochemicals, Jealott's Hill Research Station, Bracknell, Berkshire RG12 6EY, England

Received August 20, 1992

A necessary precursor to QSAR analysis is the creation of a data table containing biological activities and molecular and substituent descriptors. This in turn requires that the molecules in the data set be built in their hypothesized three-dimensional active conformation and that their atoms and substituents are labeled in a consistent manner. Using a mixture of novel and commercially available programs, we have devised and implemented a software system to facilitate these tasks. The key program, AUTONAME, enables the atom and substituent labels of a template molecule to be defined. These are then transferred to individual molecules in the QSAR data set by substructure matching. Problems due to symmetrical substructures are overcome by the introduction and use of the concept of "substituent priorities". This enables the investigator to define a property by which substituent priorities will be determined (for example, volume or electron-withdrawing ability). When labeling the substituents at two symmetrically equivalent positions on a substructure, AUTONAME will then ensure that the substituent of higher priority is always assigned the same label.

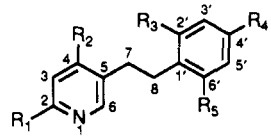
INTRODUCTION

The study of quantitative structure-activity relationships (QSARs) is an established technique for rationalizing and, in favourable circumstances, predicting the biological activities of molecules.[1] Typically, a series of related molecules is taken, all of which are presumed to bind to the same biological receptor or enzyme active site. Each molecule is described by several parameters, which characterize size, shape, electronic and physical properties, and so on. These are the "descriptor variables". Statistical methods such as multiple regression are then used to find an equation relating the measured biological activities of the molecules to the descriptor variables.

The availability of statistical programs such as SAS[2] and SYBYL QSAR[3] ensures that the statistical analyses involved in a QSAR study are easy to perform. In contrast, the construction of the data table of descriptor variables is time consuming: in our experience, a comprehensive table of descriptors for a large series of molecules—say, 300 or 400—may take several days to compile. We have, therefore, designed and implemented software that facilitates construction of tables of descriptor variables, reducing the effort required by about an order of magnitude. This system is described below.

TYPES OF DESCRIPTOR VARIABLES

Consider, as a hypothetical example, the series of molecules **1–5**.



	R_1	R_2	R_3	R_4	R_5
1	Cl	F	Cl	i-Pr	F
2	F	t-Bu	Cl	F	F
3	Me	Me	H	t-Bu	Br
4	Et	SMe	Cl	H	Et
5	H	H	H	H	H

A fundamental distinction may be made between descriptors that are properties of the whole molecule (e.g., dipole moment), those that pertain to a particular substituent (e.g., volume of R_1), and those that relate to an individual atom (e.g., partial charge on atom 2'). All three types are used in QSAR studies. Clearly, the substituent- and atom-based descriptors only have meaning if a consistent labeling scheme is used throughout, so that, for example, substituent R_1 of **1** corresponds to R_1 of **2–5**.

Descriptors also vary according to the amount of information that is required for their calculation. Thus, some descriptors can be computed (or looked up) solely from 2-D chemical connectivity information. Examples are the water/octanol partition coefficient, log P, and substituent π values.[4] Other descriptors, such as molecular volume, depend on 3-D geometry. For flexible molecules such as **1–5**, this requires the investigator to hypothesize the active conformation of the molecules (i.e., the conformation adopted at the biological binding site) and to build the 3-D structures accordingly. Finally, vector properties such as dipole moment depend not only on 3-D geometry but also on molecular orientation. Thus, their calculation demands that the molecules be overlaid on one another, so that their atomic coordinates are expressed with respect to a constant reference frame.

STEPS REQUIRED FOR CONSTRUCTION OF QSAR DATA TABLES

Most pharmaceutical and agrochemical companies maintain databases in which are stored the biological activities and 2-D chemical connectivities of all compounds synthesized in-house. The usual first step in a QSAR study is to perform one or more substructure searches, in order to retrieve the activity and connectivity information for the molecules of interest. In addition, the investigator will normally undertake molecular modeling calculations, so as to identify the active conformation of the molecules.

Starting from this point, a number of steps are required for construction of the QSAR data table:

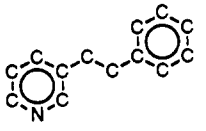
1. The raw biological data must be manipulated to give a summary activity score for each compound. This step tends to be very dependent on the exact nature of the biological test and is not discussed here.

0095-2338/93/1633-0174$04.00/0

2. Atom labels must be assigned to each molecule in a consistent manner (e.g., so that atom 1 in **1–5** is always the pyridine nitrogen).
3. Similarly, substituents must be defined and labeled consistently. In each molecule, the atoms comprising each substituent must be identified.
4. The 3-D structures of the molecules must be built in the hypothesized active conformation.
5. The molecules must be overlaid on one another.
6. The various molecule, substituent, and atom descriptors may then be computed or looked up and collated into the required data table.

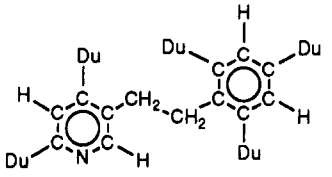
SOFTWARE FOR CONSTRUCTION OF QSAR DATA TABLES

Basic Strategy. Most QSAR studies are performed on sets of closely related molecules; indeed, some authorities argue that this is a prerequisite for a successful analysis.[5] For example, all molecules in the hypothetical series **1–5** contain the common (or *generic*) substructure **6**.



6

We have exploited this fact in developing software to facilitate construction of QSAR data tables. Thus, our software enables the user to define the atom- and substituent-labeling scheme of a *template* molecule, which comprises the generic substructure with dummy atoms representing the substituents. For example, **7** would be a suitable template for **1–5**, the five dummy atoms representing the substituents R_1–R_5.



7

Substructure matching is used to map the template onto each molecule of the QSAR data set in turn. It is then straightforward to transfer the atom labels from template to molecule. Also, the labeling of the substituents and the identities of the atoms that comprise them can be inferred from the mapping of the template dummy atoms onto the atoms of the molecule. As will be seen below, most of the remaining steps required for construction of the QSAR data table become trivial once the atom labels and substituents have been so defined.

Component Programs. Key programs in the system are as follows:

1. CONCORD.[6] Used to construct an initial 3-D structure for each molecule in the data set.
2. SYBYL.[3] Used for building templates and for performing a variety of modeling calculations.
3. AUTONAME. Used to assign atoms labels and substituent data to molecules by substructure matching onto a template.
4. PARRET. Used to retrieve substituent descriptors.

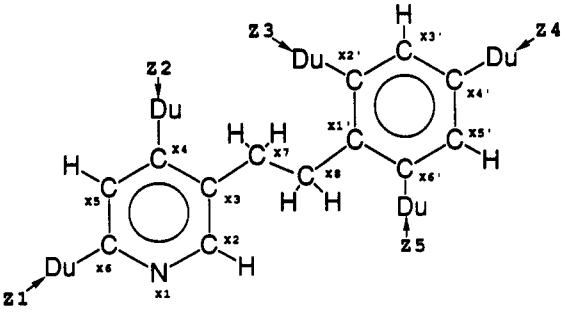


Figure 1. Template for molecules **1–5**. Du = dummy atom; Z1–Z5 = substituent labels; X1–X6′ = atom labels.

5. MAKETAB. Used to collate selected descriptor variables into a single data table.

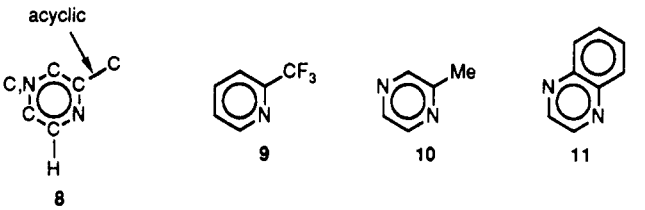
The software is implemented on a Silicon Graphics 4D/220 Power Iris, under the Irix operating system.

Generation of Preliminary 3-D Structures. Output from substructure searches of ICI Agrochemicals' chemical compound database can be written to a *Smiles* file. This completely specifies the 2-D chemical connectivity of every molecule found by the search(es), using the notation devised by Weininger.[7] From this information, the program CONCORD can generate a 3-D structure for each molecule in the file, using a rule-based technique. The structures are usually low in energy but are not necessarily the active conformation. However, the purpose of this step is merely to generate an initial 3-D structure for each molecule in the QSAR data set.

Template Building. Templates are built with the molecular graphics program SYBYL, using a short macro written in Sybyl Programming Language (SPL).[3] When building a template, the user may assign a unique label to any atom in the generic substructure. By convention, we use 2- or 3-character labels beginning with the letter X (this helps later in distinguishing meaningful atom labels from those assigned arbitrarily by SYBYL or CONCORD). Each dummy atom in the template—representing a substituent in an actual molecule—is assigned a *substituent label* (by convention, 2- or 3-characters beginning with the letter Z). Thus, Figure 1 shows a suitable template for **1–5**.

Once built, the template is saved as a SYBYL *mol2* file. The atom labels are stored in the *atom name* fields of the mol2 file, and the substituent information is stored in STATIC SET records.

Templates may contain variable atoms, and individual bonds may be designated cyclic or acyclic. For example, template **8** would match onto molecules **9** and **10** but not **11**.



8 **9** **10** **11**

Variable atom and bond cyclicity information is stored in STATIC SET records of the template mol2 file.

Use of Substructure Matching To Assign Atom Labels and Substituent Data. The program AUTONAME is used to find substructure matches between the template and individual molecules in the QSAR data set. Once a match is found, atom and substituent labels are assigned to the molecule, and the atoms comprising each substituent are identified. For example, Figure 2 shows the assignment of atom labels and

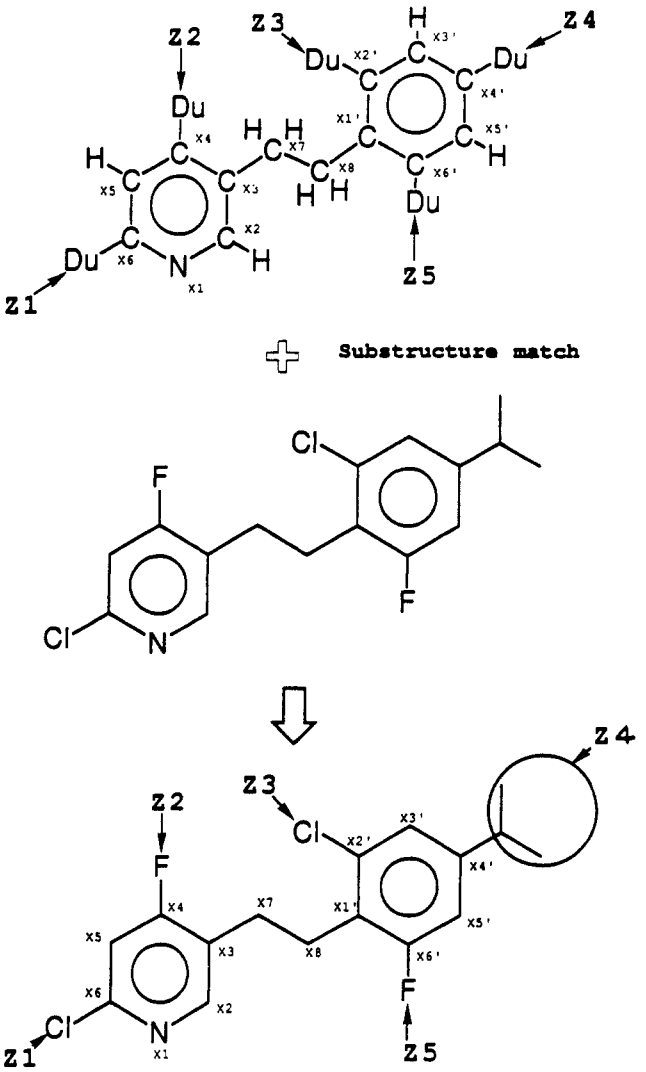


Figure 2. Matching of template (top) onto molecule **1** (middle) to give molecule **1** with atom labels and substituent data defined (bottom).

substituent data to **1** by substructure matching onto the template of Figure 1.

Once processed by AUTONAME, a molecule is written out in mol2 format; the atom labels and substituent information being stored in the atom name fields and STATIC SET records, as for the template.

Refinement of 3-D Structures; Molecular Overlays. Once their atoms have been labeled consistently, it is straightforward to manipulate the molecules of a QSAR data set into a hypothesized active conformation. For example, a single SYBYL command can be used to set the torsion angle X4-X3-X7-X8 of **1** (see Figure 2) to any user-specified value. Assuming **2–5** have also been processed by AUTONAME, the identical command can be used on them. Use of an SPL macro enables the command to be executed automatically on each molecule in turn. The molecules can then be overlaid, e.g., by a least-squares fit of X1, X2, ..., X6′ of **2–5** onto the corresponding atoms of **1**. Again, this can be done with a single SYBYL command, executed automatically on each molecule in turn.

Calculation of Theoretical Properties. Calculation of theoretical properties is also straightforward. At ICI Agrochemicals, we use a variety of programs for calculating conformational energies, molecular orbital energies, atomic partial charges, and so on. These programs have been integrated together, so that output from one can be used as input to another, without user intervention.[8] For example, it is possible to energy minimize each of a set of molecules by molecular mechanics and then perform a single-point ab initio molecular orbital calculation on the optimized geometry. These calculations will run automatically on each molecule in turn, output from the molecular mechanics step being piped into the molecular orbital program.

Table I. Example Molprop File[a]

NAME (ZB) F	Q(XB) 0.070
PI(ZB) 0.14	Q(XC) 0.007
SI(ZB) 0.54	DIPOLE 1.243
NAME(ZD) H	DIPX 0.020
PI(ZD) 0.00	DIPY 1.013
Q(XA) −0.120	DIPZ 0.720

[a] Containing names; π values of substituents ZB, ZD; Taft inductive σ of substituent ZB; partial charges on atoms XA, XB, and XC; and dipole moment magnitude and components in x, y, z directions.

Since the atoms in the molecules have been labeled consistently, it is possible to analyze the output from such calculations with ease. For example, suppose that **1–5** were used as input to a molecular orbital program: the partial charge of atom X1 could be retrieved from each output file in the knowledge that X1 refers to the same atom (the pyridine nitrogen) in all of the molecules.

Retrieval of Substituent Descriptors. Substituent descriptors (e.g., π values) are retrieved from an in-house database with the program PARRET. The required parameters are specified by using the substituent labels assigned by AUTONAME; for example, the π values and volumes of substituents Z1, Z2, and Z4 might be requested. Taking each molecule in turn, PARRET computes the formula of each substituent for which a parameter has been requested; this is straightforward, since the atoms comprising the substituent were identified by AUTONAME and the information was stored in the molecule's mol2 file. Substituent-formula and substructure matching are then used to find and retrieve the desired parameter values from the database, which is indexed on the formula for efficiency.

Collection of Descriptors into QSAR Data Tables. As descriptor variables are calculated or looked up, they are appended to *molprop files*, there being one such file for each molecule in the data set. At the end of the exercise, therefore, all of the descriptors determined for a given molecule are stored in its molprop file. The simple, keyword-based format is illustrated in Table I. Given a set of molprop files, the program MAKETAB can be used to select a subset of the descriptors and collate them into a single QSAR data table, with one row per compound in the data set. This may then be read into statistics programs such as SYBYL QSAR[3] and SAS.[2]

SUBSTRUCTURE SYMMETRY AND OTHER PROBLEMS

The preceding section describes how the AUTONAME/PARRET system facilitates construction of QSAR data tables for simple series such as **1–5**. In practice, however, a number of problems may arise. These are considered below.

Lone Pairs. Dummy atoms in the template are ignored in substructure matching, so that, for example, template **12** would be deemed to match all of **13–15**. AUTONAME will add explicit lone pairs to the connectivity specifications of **14** and **15**, and these lone pairs will be regarded as the "substituents" ZA (for **14**) and ZA,ZC (for **15**: see Figure 3).

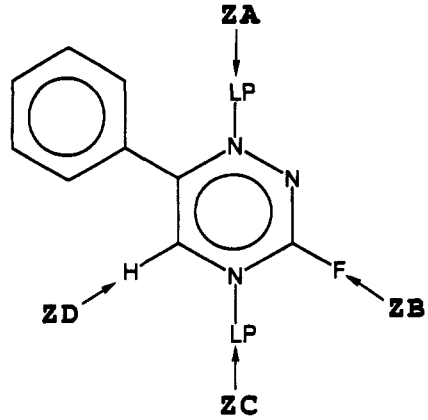
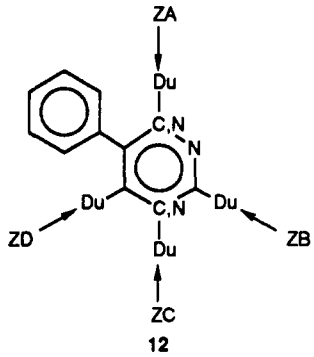
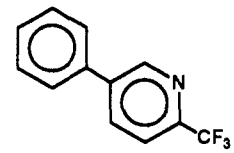


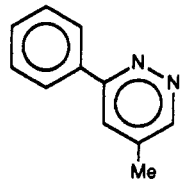
Figure 3. Result of matching template **12** onto **15**. LP = explicit lone pair.



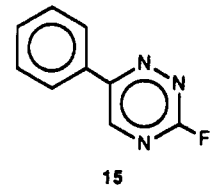
12



13

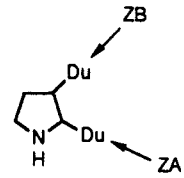


14

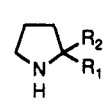


15

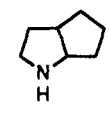
Ambiguous and Cyclic Substituents. Substructure matching of template **16** on molecules of the general formula **17** leads to ambiguity in defining substituent ZA. In this situation, the following rules are applied: if R_1 = H and R_2 = H, ZA is taken arbitrarily as either R_1 or R_2; if R_1 = H and $R_2 \neq$ H, ZA is taken as R_2; if $R_1 \neq$ H and $R_2 \neq$ H, the molecule is not processed, the user being obliged to make the decision manually.



16



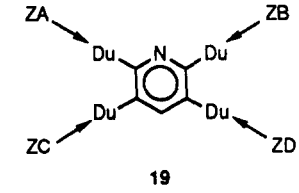
17



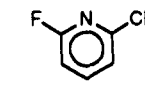
18

Molecules with cyclic substituents are also not processed, e.g., the atoms of the cyclopentane moiety of **18** would not be assigned to either substituent ZA or ZB. This rule is applied only because of the difficulty in defining meaningful substituent descriptors for such cyclic systems.

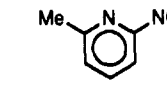
Symmetrical Substructures. Difficulties arise with symmetrical templates such as **19**, which can be substructure matched onto molecules in more than one way. This may lead to ambiguities in defining atom labels and substituent data. For example, in mapping **19** onto **20**, either the fluoro or chloro substituent could be labeled ZA; in **21**, ZA could be either the nitro or methyl group.



19



20



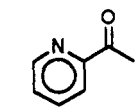
21

The correct way of labeling the substituents depends on the relative orientations of **20** and **21** when bound to a common biological receptor. If the chloro and methyl substituents of **20** and **21**, respectively, occupy the same binding pocket, they should be assigned the same substituent label (i.e., both ZA or both ZB). Unfortunately, this sort of information is almost never available when QSAR studies are performed. Consequently, the investigator will wish to explore a number of hypotheses, e.g., by labeling the *larger* substituent ZA and the smaller ZB, or by labelling the *more electronegative* ZA and the less electronegative ZB, and so on. Separate QSAR analyses based on each of these hypotheses in turn may indicate which is correct.

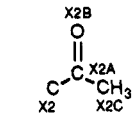
In order to facilitate this type of study, the user of AUTONAME is permitted to specify that the *priority* of substituent ZA must not be less than that of ZB. Substituent priority may be determined by any of the descriptors that can be retrieved from the substituent-parameter database, e.g., volume or π value. Thus, if volume is chosen, the largest substituent is deemed to have the highest priority, and the template will be mapped onto the molecule so that volume(ZA) $\geq$ volume(ZB). Nested priority specifications may be used, e.g., volume(ZC) $\geq$ volume(ZD) if volume(ZA) = volume(ZB).

Multiple Occurrence of Substructures. Construction of QSAR data tables with the aid of AUTONAME may sometimes require the use of several templates. For example, having assigned atom labels to the pyridine moiety of molecules with the general formula **22**, the investigator may decide to specify the conformation with respect to the pyridine ring of the acetyl substituent in the subseries **23**.

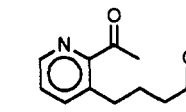
22



23



24

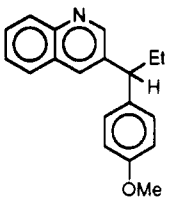


25

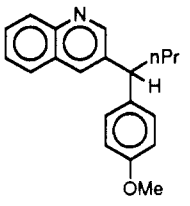
In principle, this can be achieved by performing a second AUTONAME run with the template **24**. This will label the atoms of the acetyl substituent in molecules of type **23**; it then requires the use of only a single SYBYL command to set the pyridine–acetyl torsion angle to any desired value.

This strategy may fail, however, if a molecule in the data set contains two or more acetyl groups, e.g., **25**, since the wrong group may be mapped onto the template and labeled. The problem is solved by application of the following rule within AUTONAME: if the same atom label, beginning with the letter X, is present in both the template and a molecule, then a substructure match will only be deemed valid if the two identically labeled atoms are mapped onto each other. In the above example, this ensures that only the desired acetyl group is labeled. More generally, the rule increases the usability of AUTONAME by giving the user more control over, and confidence in, how templates are matched onto molecules, thereby reducing the amount of preplanning that must be done before a series of AUTONAME runs.

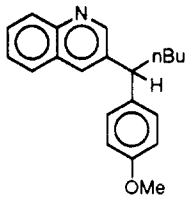
Stereochemistry. Definition of stereochemistry is absent or incomplete in most chemical databases. Consequently, the generation of 3-D structures by CONCORD cannot be done with complete reliability: sometimes, chiral centers will be built in the wrong configuration and double bonds will be made E rather than Z, or vice versa. In many cases, these problems can be solved after atom labels have been assigned systematically by AUTONAME. For example, a single SYBYL command, repeated automatically on each molecule by use of SPL, can be used to set the chirality of the optical center in **26–28** to S, provided that the asymmetric carbon atom has been assigned the same label in all of the substructures.



26



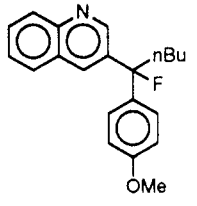
27



28

A similar method may be used to set E/Z stereochemistry.

Difficulties may still arise because of the vagaries of the Cahn–Ingold–Prelog (CIP)[9] rules. For example, the fluoro substituent in **29** reverses the priority order of the substituents, so that the R isomer is now required for optimum steric overlay on **26–28**.



29

In principle, this sort of problem could be dealt with by the same mechanism used to control matching of symmetrical substructures (see above). For example, the investigator could be allowed to override the CIP priority rules by substituent priorities based on, say, volume (this would be suitable for **26–29** above). Even more precise control could be achieved by enabling the user to define substituent priorities on the basis of atom labels (for example, substituent containing atom XA to be regarded as highest priority, that containing XB to be next highest, etc.). However, these features have not yet been implemented in our system.

Mixtures. Occasionally, the substructure search used to find the members of a QSAR data set will retrieve a database entry that is a mixture of two or more compounds. In our system, a preliminary 3-D structure of each component of the mixture will be generated by CONCORD. If only one contains the generic substructure of the series, it will be matched against the template in AUTONAME, and atom labels and substituent data assigned as usual. The other component(s) will obviously not be matched. This is the desired outcome, since the correct component can then be processed further (e.g., its substituent descriptors looked up by PARRET) while the unwanted component(s) will effectively be dropped from the data set.

If more than one component contains the generic substructure, each will be processed by AUTONAME, PARRET, etc. However, the investigator will be alerted to the presence of a mixture in the data set. The investigator must then decide what to do with it: the normal decision will be to remove all components from the analysis, since it will be unclear how any observed biological activity should be apportioned between them.

SUMMARY AND CONCLUSIONS

The system described above can reduce the labor involved in setting up a QSAR data table by an order of magnitude. Some improvements can be envisaged, notably improved control of stereochemistry, and will be implemented in due course. However, in our experience, these limitations affect only a minority of QSAR data sets.

The current implementation of AUTONAME and PARRET is tailored to our use of the SYBYL molecular graphics program (for example, mol2 files are used as the standard format for storing molecules). However, relatively little effort would be required to amend the system so that it could be used in conjunction with other molecular modeling packages.

Requests for copies of AUTONAME should be directed to Dr. R. Taylor.

ACKNOWLEDGMENT

We thank Drs. K. J. Heritage and J. A. Farrington for useful discussions and contributions to a prototype system. ICI Agrochemicals in the United Kingdom is part of Imperial Chemical Industries plc.

REFERENCES AND NOTES

(1) Fujita, T. The Extrathermodynamic Approach to Drug Design. In *Comprehensive Medicinal Chemistry*; Ramsden, C. A., Ed.; Pergamon: Oxford, 1990; Vol. 4, pp 497–560.
(2) *SAS User's Guide*; SAS Institute Inc.: Cary, NC, 1988.
(3) *SYBYL User Manual*; Tripos Associates Inc.: St. Louis, MO, 1991.
(4) Taylor, P. J. Hydrophobic Properties of Drugs. In *Comprehensive Medicinal Chemistry*; Ramsden, C. A., Ed.; Pergamon: Oxford, 1990; Vol. 4, pp 241–294.
(5) Wold, S.; Dunn, W. J., III. Multivariate Quantitative Structure–Activity Relationships (QSAR): Conditions for their Applicability. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 6–13.
(6) Rusinko, A., III; Skell, J. M.; Balducci, R.; McGarity, C. M.; Pearlman, R. S. *CONCORD, a Program for the Rapid Generation of High Quality Approximate 3-Dimensional Molecular Structures*; University of Texas: Austin, 1988.
(7) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.
(8) Taylor, R.; Mullier, G. W.; Sexton, G. J. Automation of Conformational Analysis and Other Molecular Modelling Calculations. *J. Mol. Graphics* **1992**, *10*, 152–160.
(9) IUPAC Commission on the Nomenclature of Organic Chemistry. IUPAC Tentative Rules for the Nomenclature of Organic Chemistry. *J. Org. Chem.* **1970**, *35*, 2849–2867.