Quant. Struct.-Act. Relat. *11*, 176 – 184 (1992)

# MULTICASE
# 1. A Hierarchical Computer Automated Structure Evaluation Program

**Gilles Klopman**

Chemistry Department, Case Western Reserve University, Cleveland, Ohio 44106, USA

## Abstract

A new algorithm is presented to analyze the structural features relevant to the biological activity of a set of molecules. The program called MULTICASE can, as its predecessor CASE (Computed Automated Structure Evaluator), automatically identify molecular sub-structures that have a high probability of being relevant or responsible for the observed biological activity of a learning set comprised of a mix of active and inactive molecules of diverse composition. New, untested molecules can then be submitted to the program, and an expert prediction of the potential activity of the new molecule is obtained. MULTICASE differs from CASE in a great many ways, but the major algorithmic difference is the use of Hierarchy in the selection of descriptors, leading to the concept of *Biophores* and *Modulators*.

**Key words:** Artificial intelligence, computer-aided drug design

## 1 Introduction

Structure-Activity studies are important in many areas of mechanistic and exploratory chemistry. They provide the rationale for drug development and for toxicity prediction. In a series of previous papers [1 – 4], our group has described a Computer Automated Structure Evaluation Program (CASE), based on an artificial intelligence concept whereby a new type of algorithm is used to automatically identify molecular fragments having a high probability of being relevant to the activity of molecules and assess the importance of these fragments in regard to the potency of the molecules that contain them. The CASE program does not necessarily lead to a Quantitative Structure-Activity Relationship (QSAR), although it can do so. It is a knowledge-based system designed to deal with "non-congeneric" data bases not normally amenable to treatment with QSAR type techniques. Its objective is to find structural entities that discriminate active molecules from the mass of inactive ones and its success is dependent on the validity of the fundamental hypothesis that a relationship does exist between activity and structure.

CASE also differs from other SAR techniques [5] in that it is completely automatic and learns directly from the crude data. It selects its own descriptors from the practically infinite number of possible structural assemblies and creates an ad hoc dictionary without human intervention. This is an important feature as other SAR procedures are primarily interactive between the operator and the computer, wherein the operator selects possible descriptors that are part of a fixed panel. Unless the correct descriptors are included, and the appropriate one may very well not be part of the panel, then the correlations obtained are not good. Thus, most SAR techniques will yield very good predictions for the carcinogenicity of polycyclic aromatic hydrocarbons (PAH's) when the bay region is included among the descriptors [6, 7]. The bay region, of course, has been identified as being crucial to the carcinogenicity of most PAH's [8]. However, if this region, or an equally important one, is as yet unrecognized, and is, therefore, not included among the SAR descriptors, then the prediction suffers accordingly. As we have shown elsewhere [9 – 11], CASE, when applied to a panel of PAH's will "discover" the bay region as being one of the important determinants of the carcinogenicity of this class of chemicals.

The CASE method has been successfully applied to the study of mutagens and carcinogens [5, 9 – 11, 12], as well as to the study of the structural basis of the pharmacological activity of drugs ranging from the anticonvulsant activity of benzdiazepines [13], to the activity of antimicrobial agents [14], hallucinogens [15], opiate alkaloids [16], antileukemic agents [17], beta adrenergic agents [18], inhibition of Thermolysin enzyme [19] and of sparteine monooxygenase activity [20].

The recent availability of CASE at other sites has led to studies on the specificity of hepatic azoreductase activity [21], the teratogenicity of retinoids [22, 23], Azolylpropanolone antifungal activity [24] and PAH carcinogenicity [25], others are forthcoming.

In many of these cases, satisfactory structure-activity relationships have been derived and the data bases have been organized automatically so that questions can now be asked to the program and expert answers relevant to any of the previously studied systems are available to the user practically instantaneously. Thus, while such a system allows the study of specific pharmacological properties to be made and sometimes rationalized, we now find that one of the major benefits of this "learning system" is that it keeps in memory everything it has ever learned and is capable to provide expert answers ever after.

While our experience with CASE was extremely stimulating and productive, it also became evident that a number of shortcomings still existed. These include the inability of CASE to deal with subtle geometrical differences, with synergies, and most importantly, with logical analysis and hierarchical deci-


0931-8771/92/0206-0176 $3.50+.25/0

sion making. After numerous attempts to address these problems we have now been able to resolve these issues to our satisfaction and wish to report our results in this paper. The changes are substantial and as a result, we felt that it was appropriate to call the new program by a different name so that any confusion between the two will be averted. The new program is called "MULTICASE". It includes the CASE algorithm as an option so that comparisons can be made as to the results of the two methodologies.

## 2 Methods

### *2.1 The CASE Algorithm*

The CASE methodology has been described on a number of occasions [1 – 4, 9 – 20]. Basically CASE selects its own descriptors automatically from a learning set composed of active and inactive molecules. The descriptors are easily recognizable structural fragments that are embedded in the complete molecule. The descriptors normally are linearly connected strings of atoms including, if necessary, a side chain. They can be as small as two heavy atoms (i.e. non hydrogens) but can grow as large as required to accommodate the problem at hand. They are characterized either as activating (biophore) or as inactivating (biophobe) fragments.

This ability of CASE to select biophores that are readily recognized as being part of a molecule is a major advantage of the method. Indeed, since our major aim is to elucidate the basis of the action of biologically active molecules, the identification by CASE of structural components embedded in the molecule offers a foothold that human intelligence can exploit with respect to possible structural site of metabolism or receptor binding which can lead to further hypothesis testing.

The necessary input to the CASE program consists of the chemical structures of the biologically active compounds as well as experimentally measured values of the expressed activity. Individual structures are coded in a manner appropriate for computer input by use of the KLN code [26], a linear coding technique developed to provide easy entry of molecular structures into a computer or by use of computer graphics. The KLN code of the individual molecules, or their graphical representation, together with a quantitative representation of their respective biological activities are stored within a central data file. Submission of the data base to the program initiates the CASE analysis.

CASE automatically "fragments" each molecular structure into its sub-units. Fragmentation of all the compounds generates many different molecular fragments, most of them totally unrelated to the observed activity. The program must then perform a statistical analysis to weed them out and identify those fragments that may be relevant to the observed biological activity. A binomial distribution is assumed, and any considerable deviation from a random distribution of a fragment among the active and inactive classes of molecules is indicative of potential significance to the biological activity. With this reduced set of statistically relevant fragments, the program can separate biologically active from inactive compounds.

Quantitative estimation of the potency or degree of biological expression can, if desired, be performed by a classical multivariate linear regression analysis based on the stepwise selection of a subset of descriptors. Activating and inactivating fragments, as well as calculated values [27] of the logarithm of the partition coefficient and the square of the logarithm of the partition coefficient are incorporated within a regression equation in a forward stepwise manner until no significant improvement is observed between calculated and actual values. The statistical validity of each of the variables is established by application of the F-partial statistic at the 95% confidence level [28, 29]. The coefficient of each of the fragments selected by the regression analysis is a measure of the activating/inactivating contribution made to the biological activity by the presence of the fragment.

Application of the F partial statistics is an effective way of discriminating between mere association and causality [28]. Indeed, scrambling of the learning sets, i.e., randomly reassigning biological activities to the chemicals consistently results in F-values which are insignificant, while the correct data yield values that are significant at the 95% confidence level. Failure to satisfy this requirement will cause the program to reject the analysis.

Once the computer has been "trained" with a particular data base, test compounds which were not included in the original CASE analysis can be submitted for qualitative as well as quantitative predictions. Thus, entry of an unknown chemical will result in the generation of all the inherent fragments and these will be compared to the previously identified biophores and biophobes. On the basis of the presence and/or absence of these descriptors, CASE, using Bayesian statistics, predicts probability of activity or lack thereof. In addition, and independently of the above, CASE also uses the descriptors obtained from the *ad hoc* multivariate regression analysis (QSAR) to project the likely potency of the molecule submitted for analysis. [1 – 5].

The program can learn since the data base can be continually updated with new compounds, leading to increased predictive accuracy. Our experience with CASE indicates that data bases consisting of 30 – 50 chemicals distributed among inactive, marginally active and active chemicals are required. Obviously larger data bases will yield better predictions.

### *2.2 The MULTICASE Algorithm*

CASE can handle many different classes of chemicals in a single data base providing that their activity, or lack thereof, results from a similar mechanistic path. However, the quality of the results may deteriorate if conflicting substituents effects are observed in the different classes of chemicals. Furthermore, the CASE program does not differentiate between the substructures responsible for activity and those that influence the activity.

In contrast to the CASE program, MULTICASE does not attempt to rationalize the complete data base at once. Instead, it relies on a hierarchical algorithm that breaks the learning set into logical subsets.

In order to document the differences between CASE and MULTICASE and illustrate the importance of the new methodology, we have assembled a data base of 121 miscellaneous organic molecules of which 42 are Bronsted acids (many are non-carboxylic) and 79 are not. The $pK_a$'s were obtained mostly from Lowry & Richardson's compilation [30]. This data base was selected, rather than a more relevant data base of biologically active molecules in order to illustrate our point in a situation where the relationship between structure and activity is well understood and readily interpretable. Table 1 shows the KLN [26] code of the selected acids, together with their chemical formula and their pKa.

The CASE analysis resulted in the identification of 5 major descriptors, 2 activating (biophores) and 3 deactivating (biophobes). As expected, the program selected -COOH as the main descriptor of acidity for organic molecules. A number of other fragments, of less statistical significance were also identified to explain the acidity of the non-carboxylic acids present in the data base.

Tables 2 to 5 show the results of the MULTICASE analysis for the same learning set.

## 3 Results and Discussion

### *3.1 Hierarchy*

In the CASE program, all relevant fragments are given a certain weight but no provision is made to prevent some correlation between them from occurring. Thus the Bayesian prediction of activity suffers from the possibility that correlated fragments reinforce each other. A further problem occurs in the linear regression evaluation of potencies, where fragments may have been given a substantial coefficient of activity because they act synergically or reinforce the potency generated by another fragment even though they alone would not produce activity.

Let us consider, for example, the CASE study of the acids eluded to above. The presence of a -COOH group is found to indicate high probability (>99.9%) that the molecule is an acid. However, the presence of a $-CHF_2$ in alpha to the acid group indicates a strong acid. $-CHF_2$ is thus identified as being relevant to activity and it is actually found that the QSAR contains the following terms:

$$Pk_a = 8.0 - 3.6\ n[\text{-COOH}] + \ldots . - 1.4\ n[-CHF_2] - 1.4\ n[\text{-CO-CHF}] \quad (1)$$

where n[-X] is equal to the no. of $-X$ groups contained in the molecule.

It can be seen that a molecule containing the -COOH group would have an average predicted $pK_a$ value of $8 - 3.6 = 4.4$. If that molecule also contained the $-CHF_2$ group in alpha to the CO group, the $pK_a$ value would be $8 - 3.6 - 1.4 - 1.4*2 = 0.2$, indicating a much stronger acid. However, a molecule containing the $-CHF_2$ group but not the -COOH group would still be predicted to be an acid, although we know that this would not be the case. CASE is fooled by the $-CHF_2$ fragment, a strong modulator of acid activity, and predicts the acidity of $CHF_2$-$CHF_2$ to be moderate, even though it is not an acid.

The problem is that $-CHF_2$ is not in itself an acidic functionality, it merely "modulates" the activity of the biophore "-COOH". A similar problem would be encountered in the evaluation of say the inhibition of an enzyme in which two separate substructures are required to produce the desired effect. One of these sites might be responsible for binding to the enzyme and the other for the inhibitory action. MULTICASE was designed to help solve this algorithm problem and does so by performing MULTIple CASE type analysis. In the first part, the program identifies only true biophores, i.e. those fragments found to have an unquestionable relation to activity. This is done in a hierarchical way, by selecting the first sub-structure that has the highest probability of being responsible for activity, as judged by the binomial probability that its observed distribution among active and inactive molecules is not due to chance. These molecules containing this substructure are then eliminated from the data set and the remaining compounds are then submitted to a new analysis. This procedure is then repeated until either;

a) The entire set is eliminated (i.e. enough structural features have been found to account for the activity of the entire data set), or
b) All statistically relevant sub-structures have been identified and the remaining data cannot be explained by statistically significant descriptors.

The molecules are then separated into subclasses based on the presence of each of the biophores. Now, for each subclass, a new analysis is performed to produce the modulators capable to modify the activity of each of the biophores. The system is thus hierarchical in that modulators are important only in the context of molecules containing the primary biophore. These modulators may offset the activity of the biophore by decreasing the activity, in which case they will be called biophobes or by increasing it, in which case they will be synergistic. In the latter case, it is possible that the modulator is itself an essential part of the biophore. This is particularly true if most of the compounds containing the biophore also contain the modulator. It is also to be noted that, unless the biophore is embedded in the modulator, the relative position of the biophore and its modulators is not recognized by the program. This may be important if, for example, they each are recognized by a multidentate enzyme where the two bonding sites are at a well defined relative locations.

When a new molecule is tested, the program will search its structure for the existence of a biophore. If it does not find one, the molecule will be called inactive by default. However, if it does find one, it will then search for the presence of potential modulators to arrive at a projected value for its potency.

**Table 1.** DATA BASE of organic acids.

| $pK_a$ | Code | Formula |
|---|---|---|
| (0.) | XDXX | CH(NO2)3 |
| (0.) | DYMYMYM | CH(SO2 CH3)3 |
| (0.) | DC3NC3NC3N | CH(CN)3 |
| 0.2 | FCTKFF | F3 CCOOH |
| 0.7 | GCTKGG | Cl3 CCOOH |
| 0.7 | FDTKF | F2 CHCOOH |
| 1.3 | KTTK | HOOCCOOH |
| 1.3 | GDTKG | Cl2 CHCO2H |
| 1.7 | RXTK | CH2NO2 COOH |
| 2.6 | FRTK | F CH2COOH |
| 2.8 | KTRTK | HOOC CH2 COOH |
| 2.9 | GRTK | Cl CH2CO2H |
| 2.9 | MRDGTK | CH3CH2CH Cl COOH |
| 3.4 | XC2DD2CD2D/TK | NO2-Ph-COOH |
| 3.8 | KRTK | HOCH2COOH |
| 3.9 | MC2DD2DD2C/TK | o-TOLUIC ACID |
| 4.0 | MDGRTK | CH3CH Cl CH2COOH |
| 4.0 | XRX | CH2(NO2)2 |
| 4.0 | GC2DD2CD2D/TK | Cl-Ph-COOH |
| 4.0 | BC2DD2CD2D/TK | Br-Ph-COOH |
| 4.2 | KTRRTK | HOOC(CH2)2 COOH |
| 4.2 | C2DD2DD2D)TK | Ph-COOH |
| 4.3 | KTRRRTK | HOOC(CH2)3 COOH |
| 4.4 | MC2DD2CD2D/TK | p-TOLUIC ACID |
| 4.4 | KTRRRRTK | HOOC(CH2)4 COOH |
| 4.5 | MOC2DD2CD2D/TK | MeO-Ph-COOH |
| 4.5 | KTRRRRRTK | HOOC(CH2)5 COOH |
| 4.5 | KTRRRRRRTK | HOOC(CH2)6 COOH |
| 4.5 | KTRRRRRRRTK | HOOC(CH2)7 COOH |
| 4.6 | KTRRRRRRRRTK | HOOC(CH2)8 COOH |
| 4.6 | GRRRTK | Cl (CH2)3 COOH |
| 4.8 | MTK | CH3CO2H |
| 4.8 | MRRTK | CH3(CH2)2COOH |
| 4.9 | MRTK | CH3CH2COOH |
| 4.9 | MRRRTK | CH3(CH2)3 COOH |
| 5.0 | MDMTK | (CH3)2 CH COOH |
| 5.0 | MCMMTK | (CH3)3 C COOH |
| 5.0 | MRDRMTK | (Et)2 CH COOH |
| 6.0 | DTMTMTM | CH(C=O CH3)3 |
| >8. | C2DD2DD2D)J | Ph-SH |
| >8. | MTRTM | CH3C=O CH2C=O CH3 |
| >8. | C2DD2DD2D)K | PHENOL |
| >8. | RC3NC3N | CH2(CN)2 |
| >8. | RYMYM | CH2(SO2 CH3)2 |
| >8. | MRK | CH3CH2OH |
| >8. | RD2DD2D/ | CYCLOPENTADIENE |
| >8. | XC2DD2CD2D/A | NO2-Ph-NH2 |
| >8. | DC2DD2DD2D)C2DD2DD2C)C2DD2DD2C)/ | 9-PHENYL FLUORENE |
| >8. | RD2DC2DD2DD2C)/ | INDENE |
| >8. | MTM | CH3C=O CH3 |
| >8. | RC2DD2DD2C)C2DD2DD2C)/ | FLUORENE |
| >8. | C2DD2DD2D)C3D | PHENYLACETYLENE |
| >8. | C2DRC2DD2DD2D)C2DD2DD2D)C2DD2DD2D) | 1,1,3-TRIPHENYL-PROPENE |
| >8. | C2DD2DD2D)A | Ph-NH2 |
| >8. | MC2DD2CD2D/A | Me-Ph-NH2 |
| >8. | MYM | CH3SO2 CH3 |
| >8. | DC2DD2DD2D)C2DD2DD2D)C2DD2DD2D) | (Ph)3 CH |
| >8. | C2DD2DD2D)RC2DD2DD2D) | (Ph)2 CH2 |
| >8. | MC2DD2DD2D) | TOLUENE |
| >8. | D2DD2DD2D/ | BENZENE |
| >8. | RRR/ | CYCLOPROPANE |
| >8. | MDMC2DD2DD2D) | CUMENE |
| >8. | DC2DD2DD2C)DC2DD2DD2C)/C2DD2DD2C/) | TRIPTICENE |
| >8. | DC2C/C2DD2DD2D)C2DD2DD2D)C2DD2DD2D) | TRIPHENYLCYCLO-PROPENE |
| >8. | RRRR/ | CYCLOBUTANE |
| >8. | RRRRR/ | CYCLOPENTANE |
| >8. | RRRRRR/ | CYCLOHEXANE |
| >8. | MRRM | BUTANE |
| >8. | MDMM | i-BUTANE |
| >8. | MRRRM | PENTANE |
| >8. | MRRRRM | HEXANE |
| >8. | MRRRRRM | HEPTANE |
| >8. | MRRRRRRM | OCTANE |
| >8. | MRRRRRRRM | NONANE |
| >8. | MRRRRRRRRRRRRRM | C 15 H 32 |
| >8. | MRRRRRRRRRRRRRRRRRRM | C 20 H 42 |

**Table 1.** DATA BASE of organic acids.

| $pK_a$ | Code | Formula |
|---|---|---|
| >8. | MRRTH | BUTANAL |
| >8. | MRRRTH | PENTANAL |
| >8. | MRRRRTH | HEXANAL |
| >8. | MRRRRRTH | HEPTANAL |
| >8. | MRRRRRRTH | OCTANAL |
| >8. | MRRRRRRRTH | NONANAL |
| >8. | MORM | CH3-O-CH2CH3 |
| >8. | MRORM | (CH3CH2)2 O |
| >8. | MRRORRM | (CH3CH2CH2)2 O |
| >8. | MRA | CH3CH2 NH2 |
| >8. | MRRA | CH3CH2CH2 NH2 |
| >8. | MDMA | (CH3)2CH NH2 |
| >8. | MRRRA | CH3CH2CH2CH2 NH2 |
| >8. | C2DD2DD2D)RA | Ph-CH2-NH2 |
| >8. | MRC2DD2CD2D/A | CH3CH2-Ph-NH2 |
| >8. | GC2DD2CD2D/RA | Cl-Ph-CH2-NH2 |
| >8. | R2DRA | CH2=CHCH2 NH2 |
| >8. | MRRK | CH3CH2CH2 OH |
| >8. | MRRRK | CH3(CH2)3 OH |
| >8. | MRRDKM | CH3(CH2)2CHOH CH3 |
| >8. | D2DRRR) | CYCLOPENTENE |
| >8. | D2DRRRR) | CYCLOHEXENE |
| >8. | MTRM | CH3 CO CH2CH3 |
| >8. | MRTRM | (CH3CH2)2 CO |
| >8. | MTDMM | (CH3)2 CH CO CH3 |
| >8. | MDMTDMM | (CH3)2 CH CO CH (CH3)2 |
| >8. | C2DD2DD2D)TM | Ph-CO-CH3 |
| >8. | C2DD2DD2D)RTM | Ph-CH2-CO-CH3 |
| >8. | R2DRM | CH2=CHCH2CH3 |
| >8. | MD2DRM | CH3CH=CHCH2CH3 |
| >8. | FCFFM | F3 CCH3 |
| >8. | GCGGRK | Cl3 CCH2 OH |
| >8. | MDGM | CH3CHClCH3 |
| >8. | MRDGRM | CH3CH2CHClCH2CH3 |
| >8. | ARTK | glycine |
| >8. | MDATK | alanine |
| >8. | MDMDATK | valine |
| >8. | MDMRDATK | leucine |
| >8. | KRDATK | serine |
| 3.9 | KTRDATK | aspartic acid |
| 4.2 | KTRRDATK | glutamic acid |
| >8. | ATRRDATK | glutamine |
| >8. | ARRRRDATK | lysine |
| 6.0 | C2DN2DE/RDATK | histidine |
| >8. | AC2EERRRDATK | arginine |

**Table 2.** List of MULTICASE biophores.

| Fragment | No. of FR. | In | Ma | Ac | Av. | pKA | No. |
|---|---|---|---|---|---|---|---|
| 1---2---3---4---5---6---7---8---9---10 | | | | | | | |
| CO-OH- | 45 | 8 | 0 | 37 | 4.5 | +++ | 1 |
| NO2-CH2- | 2 | 0 | 0 | 2 | 2.9 | | 2 |
| NO2-CH- | 1 | 0 | 0 | 1 | 0.0 | | 3- |
| CO-CH-CO- | 1 | 0 | 0 | 1 | 6.0 | | 4 |
| SO2-CH-SO2- | 1 | 0 | 0 | 1 | 0.0 | | 5 |
| N SC-CH-C SN- | 1 | 0 | 0 | 1 | 0.0 | | 6 |

**Table 3.** List of MODULATORS related to BIOPHORE: -COOH

| | | | | | Constant = 4.22 | |
|---|---|---|---|---|---|---|
| Fragment | No. of FR. | In | Ma | Ac | QSAR | No. |
| 1---2---3---4---5---6---7---8---9---10 | | | | | | |
| NH2-CH- | 10 | 7 | 0 | 3 | 2.97− | 1 |
| F -CH-F | 1 | 0 | 0 | 1 | −3.34 | 2 |
| CO-C-F | 1 | 0 | 0 | 1 | −2.47 | 3 |
| CO-C-Cl | 1 | 0 | 0 | 1 | −2.13 | 4 |
| CO-CH-Cl | 2 | 0 | 0 | 2 | −1.70 | 5 |
| CO-CH2-NH2 | 1 | 1 | 0 | 0 | 22.70 | 6 |

**Table 4.** MULTICASE evaluation of the acidity of $CHF_2$-COOH.

INPUT MOLECULAR CODE (or ?) : FCHF-CO-OH

FORMULA

1 2 3 5 6

F-CH -CO-OH
-F

Enter its activity (1=inactive, 2=marginal, 3=active, ?=Unknown) [?]...?

The molecule contains the biophore:
CO -OH

37 out ot the known 45 molecules (82%) containing such biophore are acids with an average pKa of 4.5 (conf.level=100%)
The following modulator(s) is/are also present:
F -CH-F Activating -3.34

The probability that this molecule is a acid is 89.9
The compound is predicted to be EXTREMELY active
(pKa=0.88)

**Table 5.** MULTICASE evaluation of the acidity of $CHF_2$-$CHF_2$.

INPUT MOLECULAR CODE (or ?) : FCHF-CHFF

FORMULA

1 2 3 4 5 6

F-CH -CH -F
-F -F

Enter its activity (1=inactive, 2=marginal, 3=active, ?=Unknown) [?]...?

The molecule does not contain any known biophore
it is therefore presumed to be INACTIVE.

**Table 6.** CASE QSAR analysis of acid database.

| No. | Molecule | Actual | Calc |
|---|---|---|---|
| 1 | CH(NO2)3 | ++++ | ++++ |
| 2 | CH(SO2 CH3)3 | ++++ | ++++ |
| 3 | CH(CN)3 | ++++ | ++++ |
| 4 | F3 CCOOH | ++++ | ++++ |
| 5 | Cl3 CCOOH | ++++ | ++++ |
| 6 | F2 CHCOOH | ++++ | ++++ |
| 7 | HOOCCOOH | ++++ | ++++ |
| 8 | Cl2 CHCO2H | ++++ | ++++ |
| 9 | CH2NO2 COOH | ++++ | ++++ |
| 10 | F CH2COOH | ++++ | ++++ |
| 11 | HOOC CH2 COOH | ++++ | ++++ |
| 12 | Cl CH2CO2H | ++++ | ++++ |
| 13 | CH3CH2CH Cl COOH | ++++ | ++++ |
| 14 | NO2-Ph-COOH | ++++ | +++ |
| 15 | HOCH2COOH | ++++ | +++ |
| 16 | o-TOLUIC ACID | ++++ | +++ |
| 17 | CH3CH Cl CH2COOH | ++++ | +++ |
| 18 | CH2(NO2)2 | ++++ | + |
| 19 | Cl-Ph-COOH | ++++ | +++ |
| 20 | Br-Ph-COOH | ++++ | +++ |
| 21 | HOOC(CH2)2 COOH | +++ | ++++ |
| 22 | Ph-COOH | +++ | +++ |
| 23 | HOOC(CH2)3 COOH | +++ | +++ |
| 24 | p-TOLUIC ACID | +++ | +++ |

**Table 6.** CASE QSAR analysis of acid database.

| No. | Molecule | Actual | Calc |
|---|---|---|---|
| 25 | HOOC(CH2)4 COOH | +++ | +++ |
| 26 | MeO-Ph-COOH | +++ | +++ |
| 27 | HOOC(CH2)5 COOH | +++ | +++ |
| 28 | HOOC(CH2)6 COOH | +++ | +++ |
| 29 | HOOC(CH2)7 COOH | +++ | +++ |
| 30 | HOOC(CH2)8 COOH | +++ | +++ |
| 31 | Cl (CH2)3 COOH | +++ | + |
| 32 | CH3CO2H | +++ | +++ |
| 33 | CH3(CH2)2COOH | +++ | ++ |
| 34 | CH3CH2COOH | +++ | +++ |
| 35 | CH3(CH2)3 COOH | +++ | + |
| 36 | (CH3)2 CH COOH | +++ | +++ |
| 37 | (CH3)3 C COOH | +++ | +++ |
| 38 | (Et)2 CH COOH | +++ | +++ |
| 39 | CH(C=O CH3)3 | ++ | ++ |
| 40 | Ph-SH | - | - |
| 41 | CH3C=O CH2C=O CH3 | - | + |
| 42 | PHENOL | - | - |
| 43 | CH2(CN)2 | - | - |
| 44 | CH2(SO2 CH3)2 | - | - |
| 45 | CH3CH2OH | - | - |
| 46 | CYCLOPENTADIENE | - | - |
| 47 | NO2-Ph-NH2 | - | - |
| 48 | 9-PHENYL FLUORENE | - | - |
| 49 | INDENE | - | - |
| 50 | CH3C=O CH3 | - | - |
| 51 | FLUORENE | - | - |
| 52 | PHENYLACETYLENE | - | - |
| 53 | 1,1,3-TRIPHENYLPROPENE | - | - |
| 54 | Ph-NH2 | - | - |
| 55 | Me-Ph-NH2 | - | - |
| 56 | CH3SO2 CH3 | - | - |
| 57 | (Ph)3 CH | - | - |
| 58 | (Ph)2 CH2 | - | - |
| 59 | TOLUENE | - | - |
| 60 | BENZENE | - | - |
| 61 | CYCLOPROPANE | - | - |
| 62 | CUMENE | - | - |
| 63 | TRIPTICENE | - | - |
| 64 | TRIPHENYLCYCLOPROPENE | - | - |
| 65 | CYCLOBUTANE | - | - |
| 66 | CYCLOPENTANE | - | - |
| 67 | CYCLOHEXANE | - | - |
| 68 | BUTANE | - | - |
| 69 | i-BUTANE | - | - |
| 70 | PENTANE | - | - |
| 71 | HEXANE | - | - |
| 72 | HEPTANE | - | - |
| 73 | OCTANE | - | - |
| 74 | NONANE | - | - |
| 75 | C 15 H 32 | - | - |
| 76 | C 20 H 42 | - | - |
| 77 | BUTANAL | - | - |
| 78 | PENTANAL | - | - |
| 79 | HEXANAL | - | - |
| 80 | HEPTANAL | - | - |
| 81 | OCTANAL | - | - |
| 82 | NONANAL | - | - |
| 83 | CH3-O-CH2CH3 | - | - |
| 84 | (CH3CH2)2 O | - | - |
| 85 | (CH3CH2CH2)2 O | - | - |
| 86 | CH3CH2 NH2 | - | - |
| 87 | CH3CH2CH2 NH2 | - | - |
| 88 | (CH3)2CH NH2 | - | - |
| 89 | CH3CH2CH2CH2 NH2 | - | - |
| 90 | Ph-CH2-NH2 | - | - |
| 91 | CH3CH2-Ph-NH2 | - | - |
| 92 | Cl-Ph-CH2-NH2 | - | - |

**Table 6.** CASE QSAR analysis of acid database.

| No. | Molecule | Actual | Calc |
|---|---|---|---|
| 93 | CH2=CHCH2 NH2 | - | - |
| 94 | CH3CH2CH2 OH | - | - |
| 95 | CH3(CH2)3 OH | - | - |
| 96 | CH3(CH2)2CHOH CH3 | - | - |
| 97 | CYCLOPENTENE | - | - |
| 98 | CYCLOHEXENE | - | - |
| 99 | CH3 CO CH2CH3 | - | - |
| 100 | (CH3CH2)2 CO | - | - |
| 101 | (CH3)2 CH CO CH3 | - | - |
| 102 | (CH3)2 CH CO CH (CH3)2 | - | - |
| 103 | Ph-CO-CH3 | - | - |
| 104 | Ph-CH2-CO-CH3 | - | - |
| 105 | CH2=CHCH2CH3 | - | - |
| 106 | CH3CH=CHCH2CH3 | - | - |
| 107 | F3 CCH3 | - | - |
| 108 | Cl3 CCH2 OH | - | - |
| 109 | CH3CHClCH3 | - | - |
| 110 | CH3CH2CHClCH2CH3 | - | - |
| 111 | glycine | - | - |
| 112 | alanine | - | - |
| 113 | valine | - | - |
| 114 | leucine | - | - |
| 115 | serine | - | - |
| 116 | aspartic acid | ++++ | ++++ |
| 117 | glutamic acid | +++ | ++ |
| 118 | glutamine | - | - |
| 119 | lysine | - | - |
| 120 | histidine | ++ | - |
| 121 | arginine | - | - |

No. of True Positives = 38 No. of False Positive = 0
No. of True Negatives = 78 No. of False Negative = 1
No. of True Marginale = 0
No. of active molecules that are predicted to be marginal = 3
No. of inactive molecules that are predicted to be marginal = 1
F(20,100,0.05)=216.87
Sensitivity = 0.9524 Specificity = 1.0000
STANDARD DEVIATION OF RESIDUALS = 4.9722
INDEX OF DETERMINATION (R-SQ) = 0.97746

**Table 7.** List of MULTICASE biophores.

| Fragment | No. of FR. | In | Ma | Ac | Av.pKA | No. |
|---|---|---|---|---|---|---|
| 1---2---3---4---5---6---7---8---9---10---------- | | | | | | |
| CO-OH- | 45 | 8 | 0 | 37 | 4.5+++ | 1 |
| NO2-CH2- | 2 | 0 | 0 | 2 | 2.9 | 2 |
| NO2-CH- | 1 | 0 | 0 | 1 | 0.0 | 3- |
| CO-CH-CO- | 1 | 0 | 0 | 1 | 6.0 | 4 |
| SO2-CH-SO2- | 1 | 0 | 0 | 1 | 0.0 | 5 |
| N SC-CH-C SN- | 1 | 0 | 0 | 1 | 0.0 | 6 |

**Table 8.** List of MULTICATORS related to BIOPHORE: -COOH

| | | | | | Constant=4.22 | |
|---|---|---|---|---|---|---|
| Fragment | No. of FR. | In | Ma | Ac | QSAR | No. |
| 1---2---3---4---5---6---7---8---9---10---------- | | | | | | |
| NH2-CH- | 10 | 7 | 0 | 3 | 2.97- | 1 |
| F-CH-F | 1 | 0 | 0 | 1 | -3.34 | 2 |
| CO-C-F | 1 | 0 | 0 | 1 | -2.47 | 3 |
| CO-C-Cl | 1 | 0 | 0 | 1 | -2.13 | 4 |
| CO-CH-Cl | 2 | 0 | 0 | 2 | -1.70 | 5 |
| CO-CH2-NH2 | 1 | 1 | 0 | 0 | 22.70 | 6 |

**Table 9.** MULTICASE evaluation of the acidity of $CHF_2$-COOH.

INPUT MOLECULAR CODE (or ?) : FCHF-CO-OH

FORMULA

1 2 3 5 6

F-CH -CO-OH
-F

Enter its activity (1=inactive, 2=marginal, 3=active, ?=Unknown) [?]...?

The molecule contains the biophore:
CO -OH

37 out of the known 45 molecules (82 %) containing such biophore are acids with an average pKa of 4.5 (conf.level=100%)
The following modulator(s) is/are also present:
F -CH-F Activating -3.34

The probability that this molecule is a acid is 89.9
The compound is predicted to be EXTREMELY active (pKa=0.88)

**Table 10.** MULTICASE evaluation of the acidity of $CHF_2$-$CHF_2$.

INPUT MOLECULAR CODE (or ?) : FCHF-CHFF

FORMULA

1 2 3 4 5 6

F-CH -CH -F
-F -F

Enter its activity (1=inactive, 2=marginal, 3=active, ?=Unknown) [?]...?

The molecule does not contain any known biophore
it is therefore presumed to be INACTIVE.

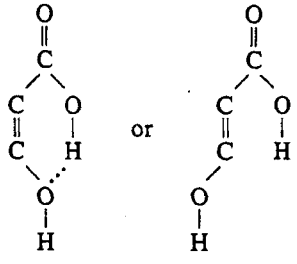
Overall, it can be said that MULTICASE deals with several "sets" of congeneric systems. The main difference between these and conventional congeneric data bases, is that the **commonality between the molecules is based on a rational evaluation of their structures rather than on an arbitrary choice of common structural features.** Table 7 shows the selected biophores and Table 8, the modulators associated with the biophore -COOH. Tables 9 and 10 illustrate MULTICASE in the predicting mode, finding again that $CHF_2$-COOH is a strong acid (Table 9). This time, however, $CHF_2$-$CHF_2$ is correctly predicted to be a non acid (Table 10).

To a certain degree, this procedure is reminiscent of methodologies based on pattern recognition [31] and principal component analysis [32] in that descriptors work in conjunction with each other rather than independently as in CASE. The methodology also bears some resemblance with the "structural alert" method proposed by Ashby [33] to identify carcinogens. Indeed, the biophores are essentially structural entities alerting to the strong possibility that the molecule is active, letting the modulators determine the quantitative activity potential. Thus the basic premise of MULTICASE is much closer to that used by the chemists than was the CASE algorithm, where all fragments relevant to activity are considered in parallel and the results obtained from the new algorithm are radically different from those obtained from the CASE algorithm.

In order to implement this new methodology, extensive modifications and considerable additions were made to the previous algorithm allowing MULTICASE to take full advantage of the graphics capabilities of the new VAXstations and provide extensive visualization of its mechanics. Nevertheless, considerable effort was made to maintain the format of the Input/Output section so as to permit easy transition from CASE to MULTICASE. A flowchart illustrating the operation of the MULTICASE program is shown in Fig. 1.

### *3.2 Incorporation of Geometry Factor*

A. **Cis-trans.** In the CASE program, the nature of the fragments were recorded by keeping track of the nature and multiplicity of the atoms that constitute the fragments and the number of hydrogens attached to them. Unfortunately, all geometrical information was lost once the fragment is taken out of context of the molecule from which it was extracted. This created a problem, particularly in systems where cis/trans information was relevant; for example, the fragment HO-C = C-CO-OH which can exist in I and in II may be linked to different chemical properties since hydrogen bonding can occur in I but not in II.




I II

In developing a suitable geometry descriptor or index, certain considerations had to be taken into account. First, the descriptor had to be compact so as not to take up too much space in the CASE data files. Second, the descriptor had to be generated before the fragment is separated from its parent compound. Third, once generated, it had to be independent of the parent compound since references to the original structure is inefficient and time consuming during the analysis of the fragments.

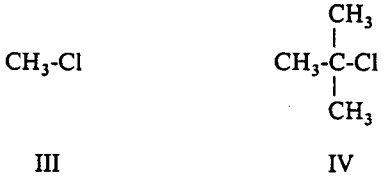
After considerable study, we found a way of describing the geometry of a fragment in a single 7-bit geometry index [34]. The index we propose is a seven bit number where each bit serves to encode the geometry of each 1,4 pair in a linear fragment. Since there are seven 1,4 pairs in a linear fragment of ten atoms, a single byte of information is sufficient to characterize the *cis-trans* configuration of a whole fragment. To generate the index, the coded structure is first decoded into a connectivity matrix. If the original code contained any geometry data, it is likewise stored in the matrix. An analysis is performed to determine the presence of rings in the structure. After the analysis, a table of all 1,4 pairs is generated where each pair is characterized as being *cis* or *trans*. By default, a pair is marked *trans*. A pair is marked *cis* if any of the following conditions is met:

a) Both atoms are originally encoded to be of *cis* configuration
b) Both atoms are the opposite pair of *cis*-encoded atoms
c) Both atoms are members of the same ring of size 8 or less
d) Both atoms are substituents of the same ring.

At the end of this process, the 1,4 table contains all pertinent information for the characterization of any fragment of any length. All that is required is to determine which atoms in the original structure make up the fragment and look up each 1,4 pair in the table of the original structure.

For example, the fragment HO-CO-C = C-OH described above has five heavy atoms in the chain. The chain contains two 1,4 pairs. In structure I, the geometry of the two 1,4 pairs are *cis* and *cis*. In compound II, the geometry of the two 1,4 pairs are *cis* and *trans*. Hence, with 1 indicating *cis* and 0 meaning *trans*, the geometry indices for each example would be 3 (11 binary) and 2 (10 binary), respectively.

B. **Fragment-Environment.** Another interesting problem consists in identifying the general environment of the fragments, and particularly the topological arrangement of the groups that surround it, e.g. the chemical behaviour of the C-Cl bond in the two structures below (III and IV) will be quite different, e.g.




III IV

For this, we turned to graph theory, and more specifically to our description of atomic graphs. In this method, we associate with each atom, a graph index which represents the geometrical complexity of the molecule (GCI), as "seen" by the atoms [35, 36]. Each fragment can now be associated with a vector obtained by averaging the vectors of all the identical fragments generated by the program. If a new molecule possesses a fragment already identified by the program as relevant, it will be reported with a "belief index" related to the vector distance between the graph value of the fragment of the molecule and the average of those used to establish the data base.

### *3.3 Expanded and Composite Fragments*

One of the major problems of "discrete descriptors" based structure-activity relationships is their inability to extrapolate results. Indeed, if a fragment is not represented significantly in the learning set, its possible importance in helping to assess the activity of a test molecule cannot be assessed, even if similar fragments have been found to be relevant. For example, one may find that $NH_2$-$CH_2$- is relevant to acidity and identified as a biophobe. However, if a test molecule contained $NH_2$-CH-, the program will fail to recognize its relevance if $NH_2$-CH- either did not appear in the molecules of the learning set or was not present in a sufficient number of cases to be of statistical significance.

This, of course, is a serious problem that so far has eluded our efforts for solution. However, it is possible to minimize its negative impact by allowing expanded and composite fragments to be considered.

**Expanded fragments.** We define an expanded fragment as a substructural entity that is similar to a bona fide biophore, but does not exist in sufficient number in the learning set to become itself a biophore.

The following rules have been drafted for identifying expanded fragments:

a) They must differ from an established biophore at only one site and only by a *minimal difference.*
b) The only acceptable *minimal differences* are:
   - A variation in the number of attached hydrogen atoms, or,
   - the replacement of a double bonded atom by another double bonded atom, e.g. C = O replaced by N = O or $NO_2$, etc...
c) The expanded fragment must have been found **only** in active molecules.

Expanded fragments, when present, are indicated by a "-" in the last column of Table 7, and are listed immediately following the biophore they approximate.

**Composite fragments** are defined as fragments that do not exist in the learning set and are constructed by an extrapolation procedure to represent structures itermediary between the biophores and their expanded analogs. For example, if in $CH_2$-COOH is found to be a biophore, and -C-COOH is one of its expanded analog, one may assume that CH-COOH should be relevant as well to activity, even though it may not have been observed among the substructures generated by the learning set. However, if it were present, but not only in active molecules, then it would not be selected as a composite fragment. Composite fragments cannot be used to make quantitative predictions but their presence may be indicative of potential activity.

### *3.4 Automatic Design*

One of the unique new features of the MULTICASE program is its ability to design molecules. Thus, using a data base of therapeutically active chemicals (e.g. beta adrenergic agents), it can use the most potent descriptor identified by QSAR to design molecules predicted to have very high activity and minimal toxicity (e.g. minimal carcinogenicity, using carcinogenic descriptors). This "autodesign feature" greatly facilitates the design of new pharmacologically important molecules.

The program normally generates a list of substructures that are relevant to activity. It is thus conceivable to use this dictionary of substructures as the basis for an expert system capable to modify submitted structures in such a way as to optimize the projected activity. Such a system would allow, for example, to have the computer *design automatically* optimized drugs.

We have implemented such an algorithm in MULTICASE, whereby a structure, submitted for evaluation can be automatically modified by the program so as to optimize its activity (or decrease it, as the case might be). This is done by searching the structure of the molecule for substructures that are similar to those listed in the dictionary, and replace them by more potent analogs. This is done in successive steps, allowing the operator to intervene and control the design path. For example, starting with toluene, the program will indicate p-nitrobenzoic acid as the optimal analog, replacing first the methyl group by a carboxyl group (biophore) and then placing a nitro group (modulator) in para to maximally increase the acidity of the resulting acid.

### *3.5 Evaluation of Metabolism*

The same methodology as that used in AUTODESIGN can be used to find the chemical entities produced by normal metabolism. The major difference between the two routines is that rather than using the dictionary generated by the program for optimal activity, an equivalent dictionary of relevant metabolic transformations is provided. We have started implementing this feature and will discuss it in more details in a forthcoming publication.

## 4 Conclusions

With MULTICASE, a number of operational problems of substructure based Structure-Activity algorithms have been addressed. The resulting MULTICASE program shares the input algorithm of CASE but is operationally considerably different from it. Our experience with the new algorithm is still preliminary but our initial evaluation seem to indicate considerable improvement in the ability of MULTICASE to predict biological activity of molecules unknown to the program [37, 38].

Furthermore, since CASE and MULTICASE share the input algorithm, all the systems we have studied in the past can be reentered via their stem file in the system and relearned by MULTICASE. We are in the process of doing this and will also report in forthcoming papers any unforeseen results that may be uncovered from this exercise.

## 5 References

[1] Klopman, G., *J. Am. Chem. Soc. 106*, 7315 (1984).
[2] Klopman, G. and Rosenkranz, H.S., *Mutation Research 126*, 227 (1984).
[3] Klopman, G. and Contreras, R., *J. Molec. Pharmacol. 27*, 86 (1984).
[4] Klopman, G., *Health and Perspectives*, J. McKinney, ed., Vol. 61, pp. 269 (1985).
[5] Frierson, M., Klopman, G. and Rosenkranz, H.S., *Environmental Mutagenesis 8*, 283 (1986).
[6] Stuper, A.J., Brugger, W.E. and Jurs, P.C., *Computer-assisted Studies of Chemical Structure and Biological Function*, Wiley, N.Y. 1979.

[7] Yuan, M. and Jurs, P.C., *Tox. Appl. Pharmacol. 52,* 294 (1980).
[8] Jerina, D.M., Lehr, R.E., Yagi, H., Hernandez, O., Dansette, P.M., Wislocki, P.G., Wood, A.W., Chang, R.L., Levin, W. and Conley, A.H., in *"In Vitro Metabolic Activation in Mutagenesis Testing";* F.J. de Serres ed., Amsterdam, Elsevier/North Holland Biomedical Press, 159 (1976).
[9] Klopman, G., Namboodiri, K. and Kalos, A., *The Molecular Basis of Cancer.* R. Rein, Editor, A.R. Liss, Inc., (1985).
[10] Rosenkranz, H.S. and Klopman, G., *Genetic Toxicology of Environmental Chemicals, Part A: Basic Principles and Mechanisms of Action,* 1986.
[11] Mitchell, C.S., Klopman, G. and Rosenkranz, H.S., *Polynuclear Aromatic Hydrocarbons: Chemistry, Characterization and Carcinogenesis, Ninth International Symposium,* M. Cooke and A.J. Dennis, eds. Batelle Press, Columbus, Ohio 1986.
[12] Klopman, G. Contreras, R., Rosenkranz, H.S. and Waters, M.D., *Mutation Research 147,* 343 (1985).
[13] Klopman, G. and Contreras, R., *Molecular Pharmacol. 27,* 86 (1984).
[14] Klopman, G., Macina, O.T., Levinson, M.E. and Rosenkranz, H. E., *Antimicrobial agents and Chemotherapy 31,* 1831 (1987).
[15] Klopman, G. and Macina, O.T., *J. Theor. Biol. 113,* 637 (1985).
[16] Klopman, G., Macina, O.T., Simon, E.J. and Hiller, J.M., *Theochem. 134,* 299 (1986).
[17] Klopman, G. and Macina, O.T., *Molecular Pharmacology 31,* 457 (1987).
[18] Klopman, G. and Kalos, A.N., *Journal of Theoretical Biology 118,* 199 (1986).
[19] Klopman, G. and Bendale, R.D., *J. Theor. Biol. 136,* 67 (1989).
[20] Klopman, G. and Venegas, R.E., *Acta Pharma. Jugosl. 36,* 189 (1986).
[21] Nesnow, S., Bergman, H., Bryant, B.J., Helton, S. and Richard, A., *J. Toxicol. and Env. Health 24,* 499 (1988).
[22] Woods, S.W. and Mass, M.J., *Environm. Mol. Mut. 11,* (suppl. 11) 114 (1988).
[23] Frierson, M.R., Mielach, F.A. and Kochar, D.M., *Fundam. & Applied Toxicology 14,* 408 (1990).
[24] Macina, O.T. and Rigby, B.S., *J. Pharm. Sc. 79,* 725 (1990).
[25] Richards, A.M. and Woo, Y., *Mutation Res. 242,* 285 (1990).
[26] Klopman, G. and Mc Gonigal, M., *J. Chem. Inf. Comp. Sci. 21,* 48 (1981).
[27] Klopman, G., Namboodiri, K. and Schochet, M., *J. Comput. Chem. 6,* 28 (1985).
[28] Klopman, G. and Kalos, A., *J. of Comput. Chemistry 6,* 492 (1985).
[29] Topliss, J.G. and Edwards, R.B., *J. Med. Chem. 22,* 1238 (1979).
[30] Lowry, T.H. and Richardson, K.S., *Mechanism and Theory in Organic Chemistry,* Harper & Row, N.Y. 1976.
[31] Stuper, A.J., Brugger, W.E. and Jurs, P.C., *ACS Symposium series 52,* 165 (1977).
[32] Wold, S. and Sjöström, M., *ACS Symposium series 52,* 243 (1977).
[33] Ashby, J. and Tennant, R.W., *Mutation Research 204,* 17 (1988).
[34] Klopman, G. and Dimayuga, M., *J. Computer-Aided Molecular Design* (in press).
[35] Klopman, G. and Raychaudhury, C., *J. Computational Chemistry 9,* 232 (1988).
[36] Klopman, G., Raychaudhury, C. and Henderson, R.V., *Math. Comput. Modelling 11,* 635 (1988).
[37] Klopman, G. and Kolossvary, I., *J. Math. Chem. 5,* 389 (1990).
[38] Klopman, G. and Rosenkranz, R.S., in press.

Received on June 25th, 1991; accepted on September 10th, 1991.

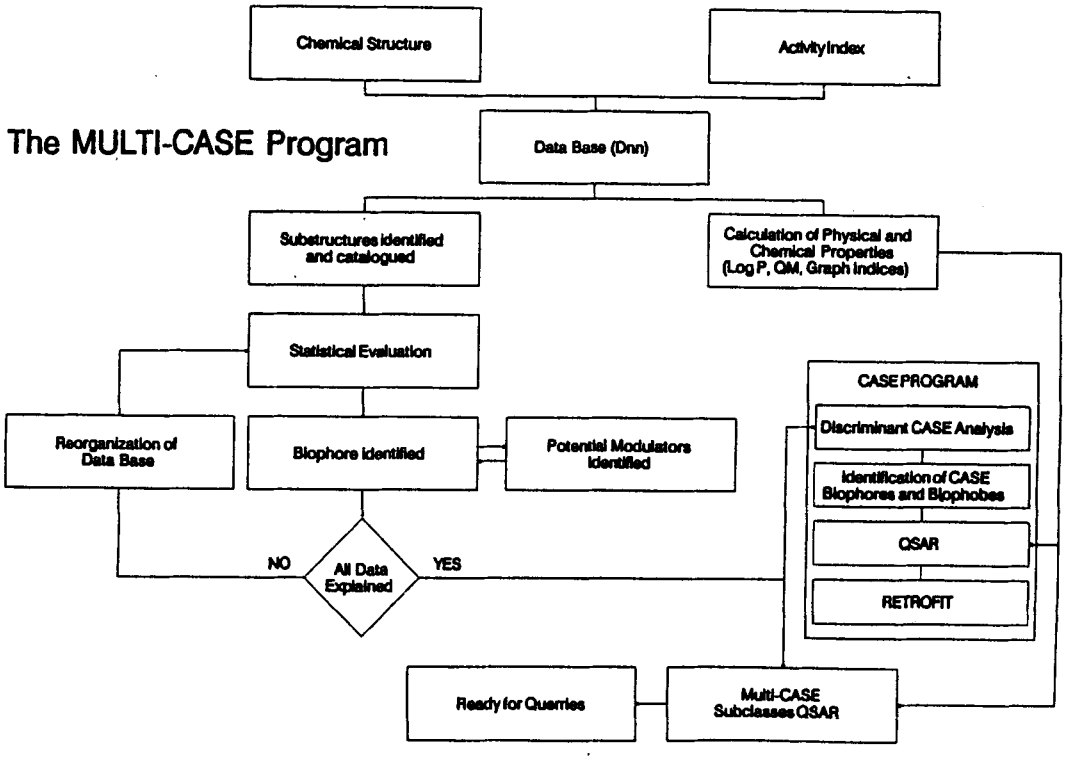



Figure 1. MULTI-CASE Program