Short-term Toxicity for Non-genotoxic Effects
Edited by P. Bourdeau *et al.*

CHAPTER 21

Computer-aided Chemical Structure-handling techniques in Structure–Activity Relationship Systems

JOSEF FRIEDRICH, WOLFGANG SCHUBERT AND IVAR UGI

2.1 INTRODUCTION

Structure–property correlation uses the occurrence of specific structural features within chemical molecules in statistical analyses. The features are correlated with physicochemical properties or biological activities. Quantitative structure–activity relationships (QSAR) have been predominantly applied to the design of drugs (see Ariens, 1971) and to the understanding of drug disposition (Lien,1981). Recently, considerable effort has been applied to the identification of chemicals with potential anticancer properties (Nasr *et al.*, 1984). The application of QSAR to the design of less toxic molecules has been recently reviewed by Hansch (1985).

21.2 STRUCTURE-HANDLING TECHNIQUES

21.2.1 General considerations

The general approach to structure-handling for SAR studies is depicted in Figure 21.1. Before any retrieval or correlation can be accomplished, the molecule file has to be processed by a feature recognition procedure. Not only structural features (such as the presence or absence of a carbonyl group) may be recognized, but also quantitative values such as topological indices or thermodynamic parameters may be included. The resulting file of structural and quantitative features will be used for further processing in retrieval or correlation.

Property estimation can be achieved by summation of substructures which contribute to a thermochemical or physicochemical property. The Hansch form of linear free energy relationship studies is a special case of the summation procedure (Kim *et al.*, 1979). The Free–Wilson method (Free and Wilson, 1964) works on the assumption that the contribution of a substituent to a given property or activity of a

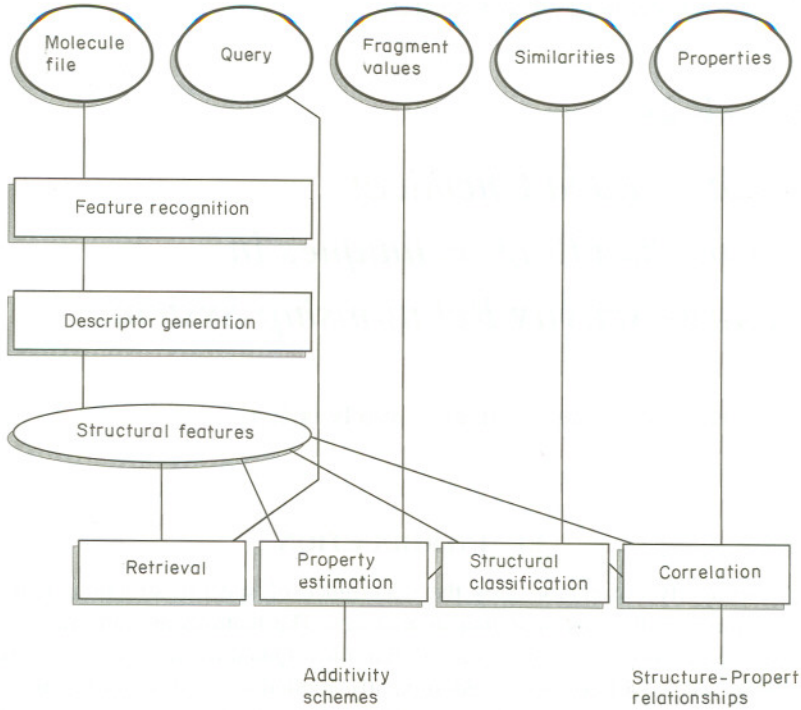

Figure 21.1 General overview of a SAR system.

molecule is independent of the contributions of other substituents at other positions in the molecule. Similarity measures, like the minimum chemical distance (Jochum *et al.*, 1982) may be used to classify structures.

21.2.2 Structural features

The simple types of structural features include basic structural units such as the total number of atoms, bonds, rings, multiple bonds, and particular atoms. Descriptors of this type do not discriminate well except in a molecule data set where all the compounds are very similar. However, in conjunction with additional structural features, they have been used in multiple regression, discriminant analyses and pattern-recognition studies.

A more sophisticated approach to defining structural features is to identify sub-structures or 'fragments' of a whole molecule. These may include, for example, rings or specific functional groups. Such variables have been widely used by manual assignment in linear discriminant analyses. Ring/functionality variables have been derived from Wiswesser line notation (WLN) for structure–toxicity analyses

(Enslein and Craig, 1979). In addition to the fragment itself, the type and relative positions of heteroatoms and substituents in six-membered ring systems were used in regression analysis of physicochemical, thermochemical and biological properties and cluster analyses.

Templates are similar to fragments but represent a 'superstructure' which is superimposed on all molecules of the data set. Free–Wilson analysis falls within this category. The template-fitting process may be done manually or automatically, derived from structures encoded in the form of connection tables.

Paths of atoms and bonds through the molecule have been used, such as a path joining two heteroatoms or chains of four to six specified atoms with bonds specified only as ring or non-ring. The most extreme approach is to take any possible path and build a molecular descriptor on the basis of autocorrelation vectors. For each property, the components of the descriptor consist of all the products of two atoms having different pathlengths between each other (Moreau and Broto, 1980). Such a descriptor has nice mathematical properties but chemical interpretation becomes difficult.

21.2.3 Feature selection

Features such as ring fragments, heteropaths, atom chains or augmented atoms may be valuable in defining pharmacophoric patterns in a single fragment. Selection of such features may be done automatically if the structures are represented by a fragment code such as WLN, or manually. Using an automated approach, the descriptors are fixed and some important substructures may not be detected. Manual selection is more flexible but is subjective and introduces bias.

Algorithmically-drived variable (feature) sets tend to represent the molecule much better than preselected sets do. For property estimation, it is essential to consider the whole molecule, so that the incremental contributions of all structural units are included. However, for correlations with physicochemical properties, a fixed set of features representing only a part of the total structure may be sufficient.

If variables overlap, problems of redundancy may arise. This could mean that some variables correlate strongly. In particular, the mixed-descriptor approach, where structural features and physicochemical parameters together constitute a descriptor, may create problems. Different types of descriptors with different weights may substitute for one another and thus create difficulties in interpretation. These problems might be eliminated by restricting the types of descriptors; by running a step-by-step analysis procedure in order to select a subset of the descriptors or to create new variables by a transformation process.

Often, less complex variables can be more effective for prediction, as demonstrated by Adamson and Bawden (1977). Sometimes, a hierarchy of descriptor types, where a higher descriptor also contains the features of any lower one, may be better (Friedrich and Ugi, 1979, 1980).

21.3 DESIGN OF A STRUCTURE–ACTIVITY CORRELATION SYSTEM

21.3.1 Design features

An ideal QSAR system would be easy to use, have a rapid response time, and would be able to handle large data sets and large molecules. It would also incorporate a procedure for the selection of structural features which is flexible, and applicable to any kind of molecule. The procedure would also create features which do not overlap or intercorrelate, are appropriately distributed through the data set and which are not given too much weight.

The two most important system features, flexible structural features and fast response time, are difficult to incorporate. Therefore, most systems in use are based on fixed feature descriptors (such as WLN) and run as batch programs. To overcome some of these difficulties, we devised an algorithm and implemented the corresponding computer program (Friedrich and Ugi, 1980). The essence of this system is the pregeneration of all conceivable substructures in a hierarchical order (Figure 21.2). Once the complete set of substructures and their interconnections through 'father–son' relations has been established, 'OR', 'AND' and 'NOT' substructure queries are performed on the stored network of substructures without the need to analyse and manipulate graphs or to reproduce structural representations out of other fixed fragment descriptions. For instance, a search for molecules containing a certain substructure (for example, C = O) would only lead to computing the corresponding substructure point in the network, entering there, and pursuing the substructure interconnections 'upwards' towards their 'fathers' (i.e. towards the unfragmented molecules of the data set). Following the network in the other direction from a subset of tagged 'active' terminal molecules leads to the substructures which are statistically the most significant for the activity in question; they have the highest ratio of active/inactive embedding molecules.

Thus we deal not only with flexible structural features covering any kind of molecule, we also avoid strong overlap (where one structure contains others) by selecting the most 'active' substructures, which are determined to have no substructures of greater activity in the hierarchy 'above' or 'below'.

It is an important feature of the design of the substructure network that each substructure is generated and stored only once, regardless of the number of molecules or larger substructures in which it occurs. Due to the fact that smaller substructures are more likely to be encountered, the introduction of a new molecule may add only a few larger substructures to the total network. Therefore, the more molecules the data set contains, the more economically the fragmentation process works.

It may happen that the size of some big molecules will lead to a disproportionately large number of large substructures requiring excessive amounts of computer space and computing time. In order to overcome problems created by these large molecules and their large molecular fragments, without relinquishing the speed and generality of the system with respect to the more widely occurring smaller

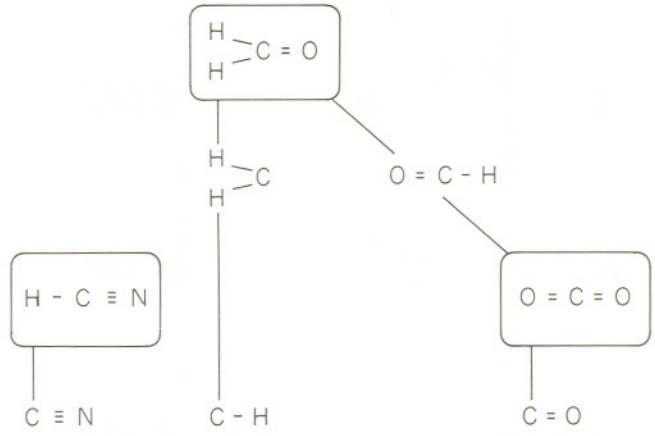

Figure 21.2 Example of a hierarchical approach for handling chemical fragments and substructures in a QSAR system.

substructures, screens of a predetermined size are generated by a fast algorithm. Here the requirements in computer time and space increase only quadratically with the chosen screen size, whereas the full substructure generation would increase with the 3rd power of the molecule's size. These screens overcome the problem of too many large substructures but also prevent the user from introducing any bias into the substructure network if manual screening procedures were used.

21.3.2 System performance

To test this system we investigated the influence of a series of substituted *o*-toluenesulphonylthioureas and *o*-toluenesulphonylureas on the level of blood sugar in rats using the data of Dove and Franke (1979). Our system automatically found the molecular fragments responsible for the effects of concern and determined the probability of incorrectly classifying the fragments. We have also tested the system using Japanese data on the accumulation of abount 100 substances in fish. These substances included polychlorinated aromatic compounds, heterocyclics, alcohols and aliphatic compounds. Figure 21.3 represents the preliminary findings of our investigation.

21.4 ACKNOWLEDGEMENTS

Development and design of this computer program was funded mainly by the Commission of the European Communities. Several versions of this program now exist. One operates in Ispra, Center of Euratom and at the University of Copenhagen, Denmark. Another version was given to the Bayer AG, Leverkusen.

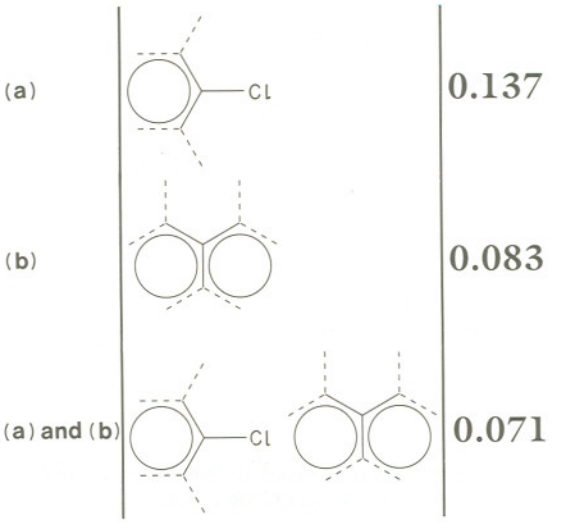

Figure 21.3 Results of a preliminary QSAR analysis of accumulation of chemicals in fish.

The Sumitomo Chemical Company received its version from the literature (i.e. from Friedrich and Ugi, 1980). The program is still under development at the Technical University of Munich.

REFERENCES

Adamson, G.W., and Bawden, D. (1977). A substructural analysis method for structure–activity correlation of heterocyclic compounds using Wiswesser Line Notation. *J. Chem. Inf. Comput. Sci.*, **17**, 164–71.

Ariens, J. (Ed.) (1971). *Drug Design*, Academic Press, New York.

Dove, S., and Franke, R. (1979). Discriminant-analytical investigation on the structure dependence of hyperglycemic and hypoglycemic activity in a series of substituted *o*-toluenesulfonylthioureas and *o*-toluenesulfonylureas. *J. Med. Chem.*, **22**, 90–95.

Enslein, K., and Craig, P.N. (1979). Status report on development of predictive models of toxicological endpoints. Genese Corp., Rochester, NY.

Free, S.M., and Wilson, J.W. (1964). A mathematical contribution to structure–activity studies. *J. Med. Chem.*, **7**, 395–9.

Friedrich, J., and Ugi, I. (1979). Substructure searchng and structure property locating by means of subgraph generation. *Match*, **6**, 201–211.

Friedrich, J., and Ugi, I. (1979). Substructure retrieval and the analysis of structure–activity relations on the basis of complete and ordered set of fragments. *J. Chem. Res.*, microfilm 1301z–80, 1401–97 and 1501–50.

Hansch, C. (1985). The QSAR paradigm in the design of less toxic molecules. *Drug Metab. Rev.*, **15**(7), 1279–94.

Jochum, C., Gasteiger, J., Ugi, I., and Dugundji, J. (1982). The principle of minimal chemical distance and the principle of minimum structure change. *Z. Naturforsch.*, **37b**, 1205–15.

Kim, K.H., Hansch, C., Fukunaga, J.Y., Steller, E.E., Jow, P.T.C., Craig, P.N., and Page, J. (1979). Quantitative structure–activity relationships in 1-aryl-2-(alkylamino)ethanol antimalerials *J. Med. Chem.*, **22**, 366–91.

Lien, E.J. (1981). Structure–activity relationships and drug disposition. *Ann. Rev. Pharmacol. Toxicol.*, **21**, 31–61.

Moreau, G., and Broto, P. (1980). The autocorrelation of the topological molecular structure: a new molecular descriptor. *Nouveau Journal de Chimie*, **4**, 359–60.

Nasr, M., Paull, K.D., and Narayanan, V.L. (1984). Computer-assisted structure–activity correlations. *Adv. Pharmacol. Chemother.*, **20**, 123–90.