Introduction to Symposium on the Development and Use of Reliable Data Bases for Quantitative Structure-Activity Relationships†

AMY RISPIN

Office of Pesticides and Toxic Substances, U.S. Environmental Protection Agency, Washington, DC 20460

Received January 10, 1981

This introduction sets forth key considerations in the development and use of data bases for prediction of quantitative structure-activity relationships in health and environmental effects. The contributions of the speakers in the symposium are introduced in this context.

INTRODUCTION

The discipline of QSAR (quantitative structure-activity relationships) is the science of predicting the biological or toxicological activity of a chemical from a knowledge of its structure. If biological end effects such as drug activity, toxicity, or pesticide potency are available for a sufficiently large set of structurally similar compounds or congeners (constituting the training set), this information can be used to predict biological properties of unknown but related compounds (the test set). Thus QSAR can be thought of as a generalization of the predictive methods used by organic chemists on the basis of properties of functional groups. In this regard, the computer can be seen as an aid to the scientist, providing him with the ability to organize and correlate large sets of end effects data by using a variety of descriptors of chemical structure. Historically the field of QSAR has focused on the prediction of pharmacological activity for drug development. Joint governmental and industrial efforts have expanded structure-activity methodologies for development of ecotoxicological agents (pesticides, animal repellants, and fish management chemicals). Currently a great deal of government, industrial, and academic interest is directed toward prediction of adverse human health effects such as cancer and other toxic effects. These efforts have led to the validation and compilation of large sets of data in the toxicological area of concern. Against this background, considerable progress has been made in the computerization of these data sets and appropriate correlation methodologies for them. Thus, it is timely to present in this symposium a collection of papers dealing with the development and use of sizeable data bases for prediction of quantitative structure-activity relationships.

CONSIDERATIONS IN THE DEVELOPMENT AND USE OF DATA BASES FOR QSAR PREDICTION OF HEALTH AND ENVIRONMENTAL EFFECTS

As the underpinning for reliable experimentation in the predictive potential of QSAR, reliable pharmacological, biological, and ecotoxicological data bases must be assembled and computerized. To be of optimum utility to the QSAR analyst, these data bases should be validated or reviewed as to data quality; the experiments should conform to appropriate and complete laboratory protocols. Data bases should also be balanced in the variety of chemical structural types and the number of positive and negative results represented. As indexes of chemical structure, appropriate clusters of physical chemical measurements must be identified along with machine coded structural details. Against this background, the performance characteristics of correlation methodologies for the prediction of structure-activity relationships can be tested.

Normally, experimental data for homologous or related compounds are generated in different laboratories in which test protocols show a range of rigor, i.e., with respect to control of animal feed, lighting, bedding, etc. In assembling such results into data bases for QSAR studies, it is important to retain experimental detail appropriate to the modeling techniques to be used; i.e., the degree of sophistication of the analysis should dictate the level of detail of the data.

Auxiliary information, such as cytoxicity data, for example, for mutagenicity results, should be retained for optimum interpretation of QSAR predictions. It is also important to the analyst to have access to critical parameters such as pH, pK, and solubility differences, for example, which qualify LC_{50} data for fish.

Biological results may be numerically imprecise and are generally not generated for use in QSAR prediction. Such data may be useful for demonstrating trends. Individual results may not be quantitatively exact. Under these circumstances, the structure-activity analyst can retain the raw data and provide confidence limits. Continuous data can be grouped as to activity or potency.

Pharmacological activity of chemicals has been successfully predicted by employing structure-activity correlation techniques to describe substrates that interact with specific biological receptor sites. Hansch analysis successfully models aqueous-lipid transport problems by using the octanol-water partition coefficient.[1,2] It uses multiple regression analysis to compare the activity of closely related series of molecules in which a few substituents are varied systematically. When a large variety of structures are represented in a data base, prediction of toxic activity requires different analytical methodologies. Discriminant analysis, cluster analysis, and other pattern recognition techniques can be applied to large diverse data sets, provided enough chemical representatives of each structural type are available (See the paper by Leftovitz et al.[3] for a comparison of different modeling programs.)

Principal components analysis can be used to isolate chemicals which act by different mechanisms. The paper by Dunn and Wold[4] will discuss the use of SIMCA, a principal components model, to classify *N*-nitroso compounds as to biological activity. SIMCA can be used to isolate chemicals which are mechanistically similar in their action.

In predicting activity in such areas as carcinogenesis and mutagenesis, the analyst is confronted with a problem which is biologically complex. Most carcinogens and mutagens must be activated to form electrophilic species which bind to DNA. Mixed-function oxidases are a family of enzymes which activate precarcinogens. The substrates for mixed-function

†Presented on April 23, 1980, as part of the symposium on Development and use of Reliable Data Bases for Quantitative Structure-Activity Relationships (QSARs) during the 14th Middle Atlantic Regional Meeting of the American Chemical Society, King of Prussia, PA.

oxidases are highly variable in size and chemical structure. Therefore, prediction of carcinogenicity may be a function–activity rather than a structure–activity problem.

The paper by Tinker[5] describes that author's adaptation of the Hodes correlation method[6] to predict mutagenic activity of chemicals. The method is suited for use with large data bases (having several thousands of molecules) and a wide variety of chemical structures. Data are classified as to substructural features contributing to similar modes of activity. The program is validated against bacterial mutagenesis testing.

For correlation with carcinogenicity of chemicals, new clusters of chemical parameters may also need to be identified and compiled in data bases for this purpose. For example, Kaufman[7] and Loew[8] are two theoretical chemists who have been deriving quantum mechanical parameters which can be used for correlations in this area.

The structure–activity analyst usually must predict biological effects in humans by using experimental results performed in laboratory animals or in vitro. Ecotoxicological predictions for genetically diverse species may have to be based on results in laboratory animals with homogenous gene pools. Species differences may account for differences in biological response. Metabolic pathways may vary qualitatively or quantitatively. For these reasons, it is valuable for the QSAR data base to couple end effect values with information about test protocols and experimental species used. The paper by Walker et al[9] concerns the development of an information retrieval system for ecotoxicological data. From testing at the United States Fish and Wildlife service, results were compiled for chemical substances hazardous to fish, wildlife, and microorganisms. In the course of testing, many of the factors (mentioned above) which affect biological response were considered. Criteria were developed for critical evaluation of the data with respect to documentation and control of test conditions, precision of methods, and sensitivity of test organisms.

Because of the considerations cited above, many industrial and governmental groups are moving toward the development of benchmark data bases to be used for QSAR experimentation. Such efforts as the GENE-TOX program, sponsored by the Environmental Protection Agency, provide mutagenicity data which are validated for use in regulatory decision making. The paper on the GENE-TOX program by Waters and Auletta[10] describes the selection and review procedures of panels which evaluated the mutagenicity of chemicals in about 25 short term assays as reported in the scientific literature. Chemicals were selected from extensive files at the Environmental Mutagenesis Information Center at Oak Ridge National Laboratories. The results of the GENE-TOX program can be used to identify appropriate test systems for particular classes of chemicals.

In carcinogenicity, the NCI bioassays[11] and the Lijinsky chronic toxicity tests on *N*-nitroso compounds[12] provide test results performed in rodents with uniform protocols. The NCI antineoplasticity data base fulfills the same purpose for antitumor tests in laboratory animals.[13]

As these and other sets of data are compiled and validated, they can be assembled into data bases with appropriate clusters of physical chemical measurements and chemical structural descriptions for use in structure–activity prediction.

The paper by Lefkovitz et al.[3] describes the assembly of many of the above sources of broad validated data into the HEEDA system. This effort is sponsored by the Environmental Protection Agency for use in the implementation of the Toxic Substances Control Act. In another paper, Weir et al.[14] describe the compilation of dose–response information for diverse biological end effects for over 430 chemicals related to coal or coal products. Data were systematically compiled from the scientific literature for biochemical, physiological, and pathological responses. A computer format is described for systematic storage and retrieval of information for use in prediction of QSAR.

REFERENCES AND NOTES

(1) Hansch, C.; Clayton, J. M. "Lipophilic Character and Biological Activity of Drugs II: The Parabolic Case", *J. Pharm. Sci.* **1973**, *62*, 1–21.

(2) Hansch, C.; Dunn, W. J., "Linear Relationships Between Lipophilic Character and Biological Activity of Drugs", *J. Pharm. Sci.* **1972**, *61*, 1–19.

(3) Lefkovitz, D.; Rispin, A.; Kulp, C.; Hill, H. "The EPA Health and Environmental Effects Data Analysis", *J. Chem. Inf. Comput. Sci.*, paper in this issue.

(4) Dunn, W. J., III; Wold, Svante "An Assessment of the Carcinogenicity of *N*-Nitroso Compounds by the SIMCA Method of Pattern Recognition", *J. Chem. Inf. Comput. Sci.*, paper in this issue.

(5) Tinker, J. "Relating Mutagenicity to Chemical Structure", *J. Chem. Inf. Comput. Sci.*, paper in this issue.

(6) Hodes, L.; Hazard, G.; Geran, R.; Richman, S. "A Statistical-Heuristic Method for Automated Selection of Drugs for Screening", *J. Med. Chem.* **1977**, 20, 469–475.

(7) Kaufman, J.; Popkie, H.; Preston, H. "Ab Initio and Nonempirical MODPOT/VRDDO Calculations on Drugs, Carcinogens, Suspected Teratogens, and Biomolecules", *Int. J. Quantum Chem., Quantum Biol. Symp.* **1978**, *5*, 201-218.

(8) Loew, G.; Phillips, J.; Wong, J.; Hjelmeland, L.; Pack, G. "Quantum Chemical Studies of the Metabolism of Polycyclic Aromatic Hydrocarbons: Bay Region Reactivity as a Criterion for Carcinogenic Potency", *Cancer Biochem. Biophys.* **1978**, 113-122.

(9) Walker, C. R.; Menzie, C. M., Bowles, W. A., Jr. "Evaluation of an Information Retrieval System for Assessment of Toxicological Effects of Chemicals on Fish, Wildlife, and Ecosystem Components", *J. Chem. Inf. Comput. Sci.*, paper in this issue.

(10) Waters, M. D.; Auletta, A. "The GENE-TOX Program: Genetic Activity Evaluation", *J. Chem. Inf. Comput. Sci.*, paper in this issue.

(11) Sontag, J.; Page, N.; Saffiotti, U. "Guidelines for Carcinogen Bioassay in Small Rodents", DHEW-NIH Publication No. 76-801, Washington, DC, 1976.

(12) Lijinsky, W. "Carcinogenic and Mutagenic N-Nitroso Compounds", *Chem. Mutagens*, **1976**, *4*, 193–217.

(13) Richman, S.; Hazard, G. F.; Kalikow, A. K. "The Drug Research and Development Chemical Information System of NCI's Developmental Therapetics Program", *ACS Symp. Ser.* **1978**, *No. 84*, Chapter 13.

(14) Weir, B. R.; Simmons, W. S.; Fan, A. M.; Livingston, D. L.; Tesche, N. S.; Walton, N. S. "Development of a Format for Abstracting Dose–Response Information from Published Studies for Use in Quantitative Structure–Activity Relationships (QSARs)", *J. Chem. Inf. Sci.*, paper in this issue.