Development of a Format for Abstracting Dose-Response Information from Published Studies for Use in Quantitative Structure-Activity Relationships (QSARs)†

B. R. WEIR,* W. S. SIMMONS, A. M. FAN, D. L. LIVINGSTON, N. S. TESCHE, and A. H. WALTON
Systems Applications, Inc., San Rafael, California 94903

Received July 31, 1980

The Toxic Substances Control Act subjects some 70 000 chemicals to regulatory action. However, empirical testing of the biological activities of this number of compounds is not feasible. An attractive alternative is the development of predictive methodology which can be used to estimate the potency of an untested compound toward a specific biological receptor. Prerequisite to such an enterprise is the highly systematic compilation of dose-response information for a wide range of biological end points and for a wide variety of molecular species. A format is described for abstracting relevant information from published studies. The format outlines the test system, experimental conditions, response analysis, exposure protocol, and results and presents the original data, all in an organized form. Regression analysis is used to estimate thresholds and potencies in the various test systems. The data may then be used to develop a predictive methodology.

INTRODUCTION

This paper describes a format we developed for abstracting biological dose-response information for use in quantitative structure-activity relationships (QSARs). The reasons for the format, the format itself, and some of its possible uses are detailed. The primary reasons for this work are twofold and concern two topics not normally thought to have any relation to each other, the Ayatollah Khomeni and the Toxic Substances Control Act. Recent and not so recent events in the Middle East have shown the United States that foreign energy supplies are not very dependable, nor are they remaining economical. Our largest domestic energy resource is coal, and, as increased utilization of coal continues, the potential environmental hazards associated with it also increase.

The Toxic Substances Control Act requires that "adequate data should be developed with respect to the effect of chemical substances and mixtures on health and the environment". Interpretation of the words "adequate data" is where the difficulty arises. Additionally, the Resource Conservation and Recovery Act, the Federal Insecticide, Fungicide and Rodenticide Act, and the Clean Air Act also require toxicological information of one kind or another.

It currently costs in the neighborhood of $500 000, requires many mice and rats, and takes up to several years to test a single compound for cancer alone. So far 431 organic compounds associated with coal and coal utilization technology and over 1000 compounds found in the air or water of the United States have been identified as environmental pollutants. With just these few number of compounds identified so far, it is still not feasible to perform live animal cancer testing on all of them, much less to test for other toxicological effects. Some other alternative to empirical testing of all the compounds must therefore be developed. We are developing a compilation of quantitative biological information for use in performing quantitative structure-activity relationships. Quantitative dose-response information has been used in the past by pharmaceutical companies to predict, by QSARs, the potency of an untested compound toward a specific biological receptor. The same approach can be used to predict biological activities for environmental pollutants.

DOSE-RESPONSE CORRELATION

For a given response, when some function of the dose is plotted against the response for a series of similar compounds, dose-response curves can be drawn. Then a standard response—for example, the ED_{50}, the effective dose that elicits a 50% response—can be extrapolated for each of the compounds, as shown in Figure 1. The ED_{50}s and the slopes of the dose-response curves can be compared to determine the relative potencies of the compounds and, through various QSARs, used to predict responses of chemicals for which no biological data exist but whose physical and chemical properties are known.

For pharmaceuticals, the responses that are being used have often been enzyme inhibition or induction or some other therapeutic activity. For environmental pollutants, however, the possible responses cover a broader range. We have divided the types of responses, for in vivo experiments, into three major categories: biochemical, physiological, and pathological responses, with a number of different subheadings under each category. For example, biochemical responses include enzyme activity and DNA effects; physiological responses include cardiovascular parameters and central nervous system effects among others; pathological responses include tumor counts and teratogenic responses among others. These subheadings will be more fully explained later. Included are responses which often have not been emphasized in classical toxicological studies. We are trying to focus attention on the more subtle, subtoxic biological effects of environmental pollutants, such as biochemical and physiological responses—those that could lead to the more classical toxicological responses, such as pathological responses.

METHODOLOGY FOR OBTAINING DATA

For over a year we have searched the literature for dose-response articles on the compounds mentioned earlier, and to ensure a complete literature search we have developed the following methodology.

(1) Computer Searching
  CHEMNAME and CHEMLINE
  TOXLINE, TOXBACK, and BIOSIS PREVIEWS
(2) Manual Searching
  Review all prints from computer searches
  Retrieve articles
  Review articles and their references
  Write to authors for data

The compound name is first entered into two chemical dictionary files, CHEMNAME and CHEMLINE, to obtain the Chemical Abstracts Service registry number (CAS number) and all the possible synonyms, which are then entered into the toxicology files TOXLINE, TOXBACK, and BIOSIS PREVIEWS. These data bases contain citation and abstract information on articles published since 1969. By combining the names, synonyms, and CAS numbers with a stored com-

† 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.

0095-2338/81/1621-0014$01.25/0

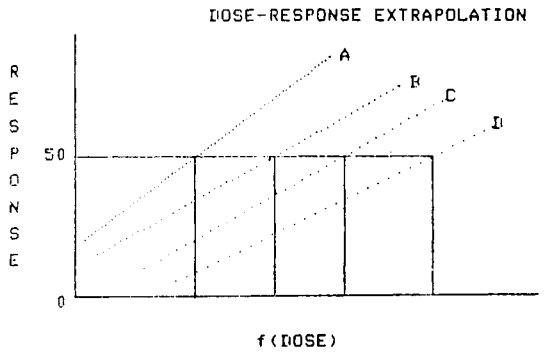

Figure 1. Extrapolation of the ED_{50}, the effective dose that elicits a 50% response, for a series of compounds, A–D.

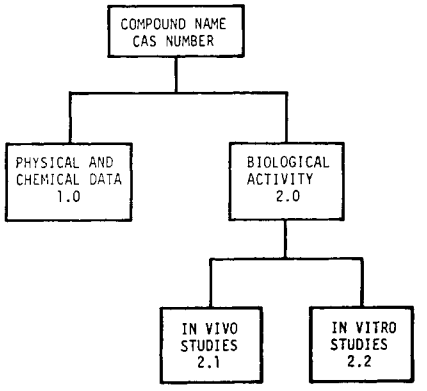

Figure 2. Format for filing of information for specific compounds. First division: Physical and chemical data divided from biological activity, which is further divided into in vivo and in vitro studies. Assigned outline numbers are given beneath each descriptor.

puter search, which we developed especially for this project, we obtain "prints", or abstracts of all articles that mention the compound of interest and any of the keywords from our computer search. We review the prints to determine which articles could contain dose–response information. Articles are retrieved by photocopying them from various libraries in the San Francisco Bay area. The articles are reviewed and the references in them scanned for other, older articles that might have been overlooked. Often when the dose–response data are presented only graphically, we have written to the authors themselves, who frequently supply us with their original raw data. Once the article has been determined to be of interest, we profile it into our format for entry into our data base. The format was developed after the completion of the entire search process for 230 organic compounds found in the ambient atmospheres of urban areas of the United States. It is designed to systematically categorize biological experiments by presenting, in a computer-readable form, all the essential information about a biological experiment. While this is no substitute for reading the original article, it does give some means for comparison of different types of studies, and it can be stored and searched by a computer.

FORMAT FOR CATEGORIZING DATA

The format contains physical and chemical data in an open registry format and toxicological studies divided into in vivo and in vitro studies, which then branch out from there (Figure 2). The numbers beneath each descriptor are outline numbers which enable one to keep track of where everything fits in the hierarchy. The in vivo side (Figure 3) is discussed in detail below. The in vitro section has similar, but not identical, subcategories.

After the study is determined to be in vivo, certain information is abstracted from it for entry into our format. We

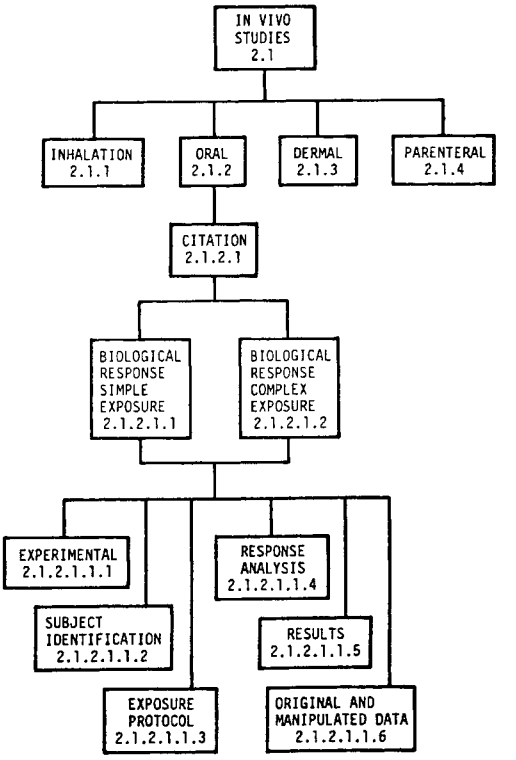

Figure 3. Format for categorizing in vivo studies, with assigned outline numbers beneath each descriptor.

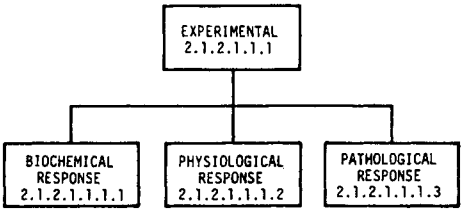

Figure 4. Major categories of experimental data, with assigned outline numbers beneath each descriptor.

first divide the study by the route of exposure used. This discussion concerns the subdivisions under oral exposure, but inhalation, dermal, and parenteral are identical save for the third number in the outline. The complete citation and whether the test animals were exposed to a single compound (a simple exposure) or two or more different compounds (a complex exposure) are included. At this point the six main subdivisions of the study are branched. They are the experimental, subject identification, exposure protocol, response analysis, results, and original and manipulated data categories. Each is detailed below in turn. It should be kept in mind that these sections are all equivalent in the hierarchy.

(1) Experimental Category. The experimental objectives are divided into three basic types of responses, as mentioned earlier—biochemical, physiological, and pathological (Figure 4). Only the biochemical responses will be discussed.

As can be seen in Figure 5, there are a great variety of responses possible. We do not claim to have categorized every possible type, just the most common ones. Room is left for additions by use of the box labeled "other" on the far right. It is apparent that the numbers are necessary to maintain distinctions in the hierarchy. For example, for sister chromatid exchange, the number 2.1.2.1.1.1.1.5 means as follows: 2—biological activity, 1—in vivo studies, 2—oral exposure, 1—first citation for this compound, 1—simple exposure, 1—experimental section, 1—biochemical response, and 5—sister chromatid exchange.

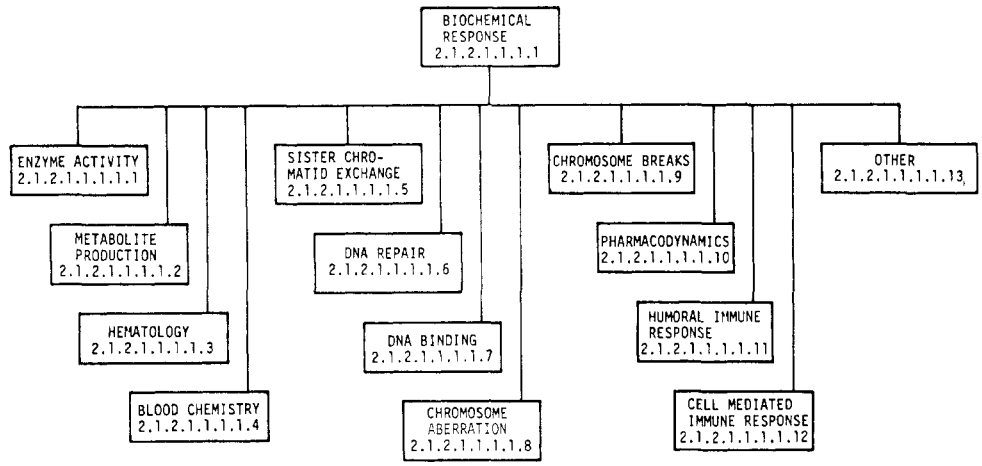

Figure 5. Format for categorizing biochemical response, with assigned outline numbers beneath each descriptor.

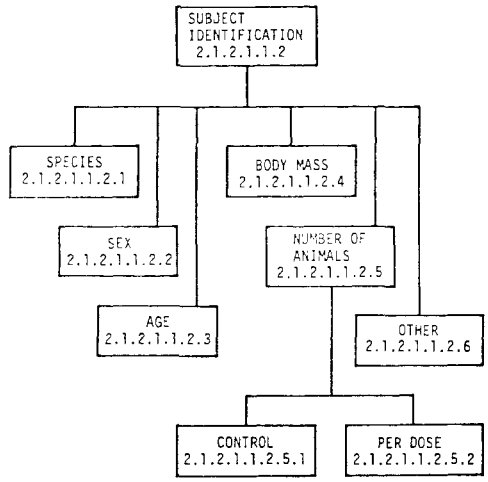

Figure 6. Format for categorizing subject identification, with assigned outline numbers beneath each descriptor.

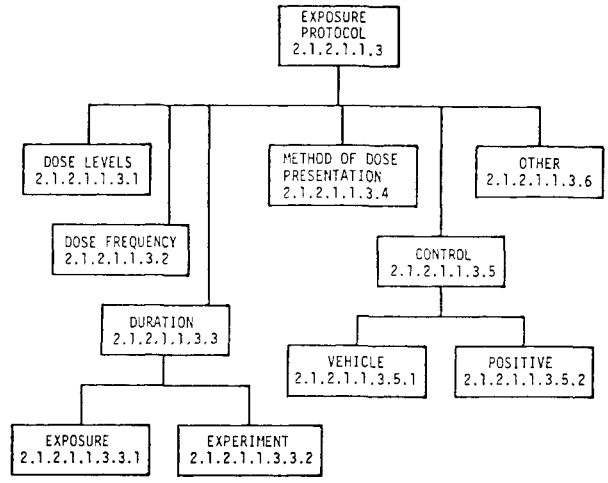

Figure 7. Format for categorizing exposure protocol, with assigned outline numbers beneath each descriptor.

There can be as many entries as necessary in these sections. For example, a study may measure an enzyme activity, a physiological response such as blood pressure, and pathological findings. All are entered into the experimental section.

(2) Subject Identification Category. The next major category, shown in Figure 6, is subject identification, which is necessary so comparisons are not made between, say, rats and sheep. The subject identification category is equivalent to the experimental section in the hierarchy. The categories include the species, sex, age, body mass of the animals, and the number of animals, both control and per dose. This latter subcategory should be emphasized. Obviously a study with only a few animals per dose is necessarily imprecise. How imprecise depends on a number of factors, but precision is important when it comes to evaluating the results. We try to obtain all the information indicated by these subcategories, but if it is not published, then we are forced to leave some sections blank.

(3) Exposure Protocol Category. The exposure protocol (Figure 7) details the specifics of exactly what dose levels were used, the frequency of dosing, the duration of both the experiment and the exposure, the exact method of dose presentation (for example, in an oral exposure, by gavage or in the diet), the controls, both vehicle and positive, and, of course, the subcategory other, where anything unusual about the exposure protocol is added.

(4) Response Analysis Category. The response analysis category (Figure 8) is equivalent to the experimental, subject identification, and the exposure protocol categories, already described, and the results and the original and manipulated data categories, which are discussed below. The response

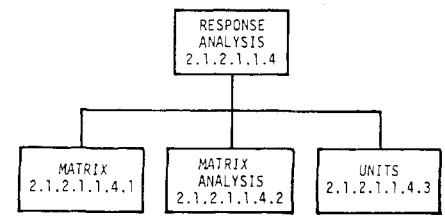

Figure 8. Format for categorizing how a response was quantitated, with assigned outline numbers beneath each descriptor.

analysis is used to describe exactly how the response was quantitated, be it spectrophotometrically, visually, or otherwise. It also details the units used for the quantitation and the matrix that the response was measured in, for example, the liver or kidney.

(5) Results Category. The results category is divided into two main sections, those results that are statistically significant and those that are not (Figure 9). The level of statistical significance that we have chosen is $P < 0.05$, commonly referred to as a 95% confidence level. The statistical tests were performed by the authors of the original articles. The nonstatistically significant results are all too often ignored, even though it is just as important to know that a compound is not teratogenic or carcinogenic as it is to know that one is. Under the statistically significant response we have many equivalent subdivisions entitled "type of response", of which only three are listed in Figure 9, but the number can be as large or as small as necessary. The exact response noted is listed in these subdivisions, such as ATPase induction, SGPT inhibition, or

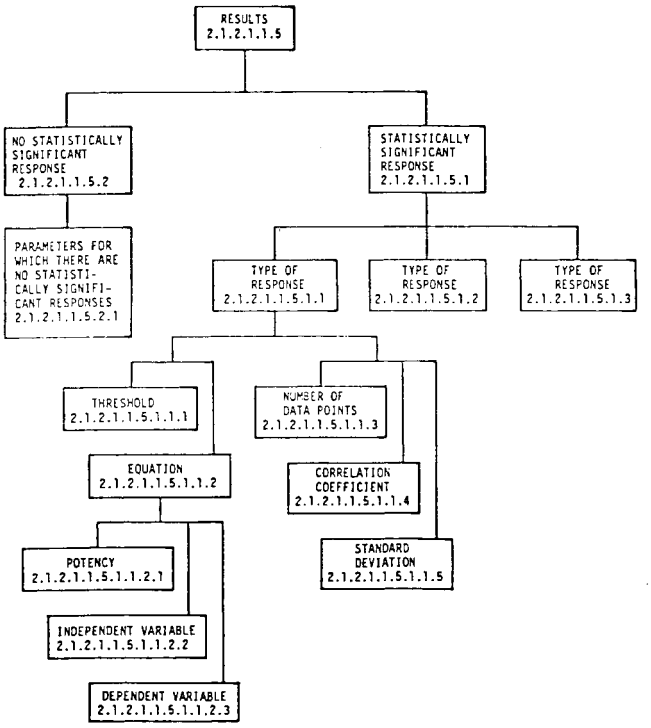

Figure 9. Format for subdividing results into those that are statistically significant and those that are not statistically significant. The numbers beneath each descriptor category are assigned outline numbers.

whatever exact response was quantitated. In this manner we are able to include, for a single study, all the different types of responses noted.

The data were manipulated to ensure a percentage increase or decrease of response relative to the control, and a linear regression analysis was performed to obtain an equation which best fits the data. The threshold mentioned in Figure 9 is the X intercept of some function of the dose vs. a function of the response or "the lowest dose above which there is an observable response" for this test system, given these data. The potency was taken to mean the slope of the dose–response curve for comparison purposes. The independent variable is a function of the dose, and the dependent variable is a function of the response. The equation is given, and certain basic statistical parameters for the regression are also included.

The data were manipulated to give the dose in molar units, since compounds react in the body on a molecular basis and any comparison of different compounds through QSAR must also be on a molecular basis. The logarithm of the dose was then taken as shown in Table I. The log dose is not always used in the regression, but it is commonly the function best suited for in vivo studies. The responses were manipulated so that the data were relative to the control, which is then taken as 0. We present the data so that if a user does not like our analysis, he can perform his own.

CONCLUSIONS

The experimental, subject identification, exposure protocol, response analysis, results, and original and manipulated data along with all the divisions above them (Figure 3) can be searched and inventoried by using our data base management system. We can also inventory all the information below these main subheadings. The software for the data base management system has been designed and is now being written; we expect to have the system operational sometime this summer.

At present we have identified over 1000 environmental pollutants, 431 of them related to coal or coal processes. We have performed the computer search on all of them and received 78 001 prints, of which 41 107 have been manually screened. Approximately 1900 articles have been retrieved and over 155 profiles prepared in preparation for entry into the computer.

Table I. Original and Manipulated Data (2.1.2.1.1.5). Effect of Pyridine on Blood Urea in Anesthetized Dogs[a]

dose, mg/kg	dose, nmol/kg	log dose	blood urea, mg %	percent increase
0	0		29	0
88	1.11	0.045	40	38
176	2.23	0.347	48.5	67
440	5.56	0.745	57	97
660	8.34	0.921	62	114
880	11.1	1.046	64	121

[a] Source: Venkatakrishna-Bhatt, H.; Shah, M. P.; Kashyap, S. K. "Toxicological Effects of Intravenous Administration of Pyridine in Anaesthetized Dogs", *Toxicology* 1975, *4*, 165-170.

This data base can then be used to

(1) establish the present body of knowledge concerning the compounds of interest, with regards to their dose–response biological activities;

(2) identify those large gaps where no dose–response information exists, although updating of the data base is simple;

(3) use in QSARs; some of the more common methodologies now used are represented by the following:

$$\log 1/C = a \log P + b \quad (1)$$

$$\log 1/C = a \log P + b (\log P)^2 + c \quad (2)$$

$$\log 1/C = a \log P - b \log (\beta P + 1) + c \quad (3)$$

Drs. Hansch and Leo of Pomona College, who developed the linear and parabolic equations [eq 1 and 2],[1-3] are working with us to perfect a predictive methodology for toxicological effects. The bilinear relationship of Drs. Higuichi and Davis[4] [eq 3] is also used, as is the molecular descriptors approach.[5] Obviously, any number of possible approaches can be taken to try to relate chemical structure to biological activity.

Overall, we have developed a format for accurately and concisely describing, in an organized computer-readable fashion, biological studies pertaining to the biological effects of chemical compounds. Provided the data are available, the data base can be used to attempt to predict specific biological activities for compounds for which no biological data are available but whose physical and chemical properties are known. Hopefully, as more data become available, our predictions can become more accurate and we can begin to predict for a broader range of compounds.

DISCUSSION

Question. I wonder what is the potential of your collection for establishing a taxonomy of toxic effects on the mechanistic level?

Answer. We have not looked at it from the mechanistic level, although it could be done. We have been trying to merely accumulate the data and then look at it once it was all accumulated on the compounds of interest. We said we had 78 001 prints on over 1000 compounds. Actually that is misleading. Most of those prints are on several well-studied compounds such as carbon tetrachloride, which has quite a bit of information, and benzo[*a*]pyrene, which has a large amount of information also. Most of the compounds that we are interested in have no biological data in the literature, or else what there is, is not necessarily what we are looking for. We have not concentrated on mechanistic data because we have not had the resources available to do so; we are concentrating more on the prediction capability right now, but mechanistic derivations could be made.

ACKNOWLEDGMENT

This work was performed for the Electric Power Research Institute under Contract No. RP 1643-1. The Project Officer was Dr. J. McCarroll.

REFERENCES AND NOTES

(1) Hansch, C.; Leo, A. J. "Substituent Constants for Correlation Analysis in Chemistry and Biology"; Wiley: New York, 1979.
(2) Hansch, C.; Clayton, J. M. "Lipophilic Character and Biological Activity of Drugs II: The Parabolic Case", *J. Pharm. Sci.* **1973**, *62*, 1-21.
(3) Hansch, C.; Dunn, W. J. "Linear Relationships Between Lipophilic Character and Biological Activity of Drugs", *J. Pharm. Sci.* **1972**, *61*, 1-19.
(4) Higuichi, T.; Davis, S. S. "Thermodynamic Analysis of Structure-Activity Relationships of Drugs: Prediction of Optimal Structure", *J. Pharm. Sci.* **1970**, *59*, 1376-1383.
(5) Stuper, A. J.; Brugger, W. E.; Jurs, P. C. "Computer Assisted Studies of Chemical Structure and Biological Function"; Wiley: New York, 1979.

EPA Health and Environmental Effects Data Analysis System†

DAVID LEFKOVITZ*
University of Pennsylvania, Philadelphia, Pennsylvania 19104

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

CAROL KULP and HELEN HILL
University of Pennsylvania, Philadelphia, Pennsylvania 19104

Received October 9, 1980

This paper discusses the development of a system to organize, store, retrieve, and correlate data pertaining to chemicals and their biological and environmental effects. The particular problems of data identification, acquisition, classification, and automation are discussed in relation to existing data sources and methods of data collection and analysis. The problems of computer software development are also addressed, and a design overview of the system is presented.

INTRODUCTION

The Office of Toxic Substances (OTS) of the U.S. Environmental Protection Agency (EPA) is charged with making regulatory decisions under the Toxic Substances Control Act (TSCA) concerning the 43 000 commercial chemicals listed in the TSCA Inventory.[1] Under various sections of the TSCA, chemicals in the inventory must be ranked for regulatory concern or selected for testing. In the course of assessing the toxicological hazard of chemicals for regulatory purposes, the computer can be used as a tool by the skilled scientist. For this reason, the Office of Pesticides and Toxic Substances (OPTS) is developing the Health and Environmental Effects Data Analysis (HEEDA) system. In providing for structure–activity prediction in toxicology, the HEEDA system contains validated or reviewed toxicological data that can be correlated statistically with structural features of chemicals. The regulatory scientists in the OPTS plan to use the techniques of quantitative structure–activity relationships (QSARs) to focus attention within the agency on chemicals of concern. Since QSARs in toxicology represent a science in its infancy, the HEEDA system will be used to assess the limitations as well as the scope of QSAR prediction.

In the creation of a software system for the processing of biological and chemical structural information, the initial phase of development must focus on data acquisition and organization (Figure 1). In this initial phase, appropriate sources of validated data must be identified for inclusion in the data base. The data must be classified biologically and chemically for ease of organization and retrieval by the analyst.

† Presented on April 23, 1980, as a part of the Symposium on Devlopment 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.

In the data access phase of development, the system is designed and programmed to provide a vehicle for information retrieval. The ability to perform substructure searches is an essential feature of data organization for retrieval of information in the chemically oriented data base. As a node in the Chemical Substances Information Network (CSIN),[2] the HEEDA system will acquire its substructure search capabilities from the CSIN Chemical Structure and Nomenclature System (CSNS).[3]

The final phase of system development is one of data correlation methodologies. HEEDA will employ two mechanisms for data correlation. The first is that of report generation to provide graphical and visual display of information in a variety of formats. The second method of data correlation is by means of mathematical modeling techniques from the discipline of QSARs.

The development of the HEEDA system has been directed toward the creation of a computer environment that contains the necessary components for structure–activity experimentation and prediction in areas of regulatory concern. At the heart of the HEEDA system is a collection of standardized, reviewed data that can be subjected to various statistical methods to correlate biological end effects with structural features. From these biological data, sets of chemicals can be assembled that are well characterized with respect to the biological effect of concern. The Office of Toxic Substances of EPA is organizing and validating many such data sets. The potential hazard of uncharacterized compounds will be assessed by comparison with the data in the training sets. Authenticated training sets can be used with different correlative techniques to test the validity of the statistical models. In the context of reliable data for training sets, different chemical–structural descriptors can be tested for their usefulness in structure–activity prediction.

0095-2338/81/1621-0018$01.25/0