Computerized Management of Structure–Activity Data. I. Multivariate Analysis of Biological Data

CORWIN HANSCH, A. LEO,* and DAVID ELKINS†

Department of Chemistry, Pomona College, Claremont, California 91711

The three articles in this series describe a system for the management and analysis of chemical structure–biological activity data wherein the chemical structures, encoded in Wiswesser Line Notation (WLN), are checked for accuracy, stored, and retrieved by computer. The first paper deals with two basic problems: (1) how to unify the description and the activity of biosystems without loss of desired flexibility, and (2) how to structure the data files.

It has become painfully apparent in recent years that even the most simple documentation of biological activity experiments must rely heavily on large-capacity computers which store only very condensed abstracts. To keep up with the world-wide flow of information from laboratories synthesizing and testing bactericides, fungicides, herbicides, insecticides, hormones, antimetabolites, pheromones, etc., organizations such as *Chemical Abstracts,* the National Library of Medicine, and *Excerpta Medica* now store and transmit much of the abstracted information on magnetic tape. These files are known respectively as "Chemical-Biological Activities," "Medlars," and "Drugdoc." Taken together with the "limited access" files of the National Cancer Institute, Walter Reed Hospital (antimalarials), and the individual pharmaceutical manufacturers, they represent a vital, almost herculean, effort to make available to the drug designer all of the available data bearing on his problem.

Ideally, this store of information should provide a valuable overview from which it might be possible to discern from the endless variety of biological responses a few broad unifying principles; but, at present, it can only serve to alert the drug designer to the full literature reports of investigations in certain selected areas, and still leaves him faced with a dilemma: specify a broad information "profile" and be inundated with a bibliography too large to digest, or choose a narrow profile and miss a key experiment because its relevance in more than one area was not apparent.

While it is true that our knowledge of pharmacology at the molecular level is so meager that the search for unifying principles may seem hopelessly premature, nevertheless the cost of drug synthesis and testing is rising so rapidly that dependence on trial and error or fortuitous by-products cannot long continue to be economic.[1] The task would indeed appear hopeless unless we are willing to delegate to a powerful computer some of the functions presently reserved for the mind of the chemist. The computer, using virtually limitless memory and rapid recall, coupled with the ability to manipulate the data in a variety of simultaneous mathematical operations, can "pre-digest" an otherwise unmanageable mass of data, reducing it to a form which the human mind can grasp, hopefully to discern some of the unifying concepts which lie therein.

It is axiomatic that in order to make structure–activity relationships quantitative, the chemist must decide upon the most significant measure of both the observed *perturbation* of the biological system and also the best measure of the *structural changes* in the perturbing agent. Since there is presently a wide disparity of opinion on what is "ideal" in both of these measurements, it is not surprising that much of the vast store of highly condensed information pouring into our data banks has so little common ground upon which to build rational structure–activity relationships.

The fundamental plan upon which the present management-analysis system is based can be represented by the equations[2]

$$\text{system}_i + (\text{drug-X}_j) \rightarrow \text{perturbation}_{ij} \quad (1)$$

$$\text{perturbation}_{ij} = \text{f(parameters of -X}_j) \quad (2)$$

In eq 1, the "system" can be any biochemical entity with a degree of complexity varying from a purified enzyme to an organelle to an intact plant or animal. The term "drug-X" usually refers to a related set of compounds such as alcohols, penicillins, barbiturates, etc., but may refer to a miscellaneous group of molecules which apparently produce a like perturbation on the system being studied. "Perturbation" refers to the action between the "system" and the "drug" and must be expressed in quantitative terms such as the concentration necessary to produce a given degree of muscle contraction. The "parameters," which are a function of the perturbation, may be derived *de novo* based on each structural variation in the set,[3-5] they may be derived from physicochemical measurements and applied in extrathermodynamic fashion,[2,6] or they may be calculated as, for example, the electron densities from molecular orbital theory,[7] or they can be a combination of the three. The function f may be linear and expressed as: $aP + c$; or it may be of a higher order: $bP^2 + aP + c$. Since a useful postulate assumes the perturbation can be expressed in terms of either a reaction rate or an equilibrium constant, both of which are free-energy related, the activity and the substituent parameters are expressed in log terms to correspond to the Gibbs expression: $\Delta F = -RT \ln K$.

It is disturbing to some people to consider that the observed response in a large, complex organism such as man is being ascribed to a perturbation in, for instance, a single enzyme system among the thousands present. They would prefer eq 2 to be considered as the sum of many reactions. Of course, it often happens that a chemical will elicit several separate but conflicting responses that tend to obscure the specific mechanisms involved. But the really effective drugs generally have one or both of the following characteristics: first, they possess a pharmacophore *unusually* well suited to interact with a receptor site on a specific enzyme or cofactor; second, the target enzyme or cofactor acts through some amplification process. Probable examples are: monoamine oxidase inhibitors which increase brain levels of biogenic amines which in turn elicit an antidepres-

* To whom correspondence should be addressed.

† Present address: G. D. Searle & Co., Box 5110, Chicago, Ill. 60680.

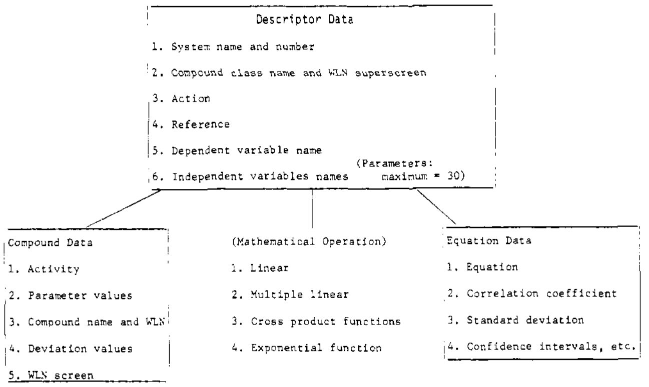


Figure 1. Contents of a "Set."

sant response, or the activation of adenylcyclase which increases the concentration of 3′,5′-cAMP which in turn activates phosphokinase and increases the active phosphorylase *b* level.[8] So, even when a chain of events is involved, it makes sense to look for a rate-controlling step which can be studied by applying eq 2.

The long-range objective of this effort is to provide a system which can produce highly relevant equations of the form of eq 2 and which, therefore, spotlight the requirements for maximum drug response in intact organisms. Additionally, these equations can be compared with those rationalizing the action of many simpler systems, the selection of one of which can be critical in elaborating the mechanism of action. Thus, ideally, the system can be helpful not only in optimizing the pharmacophores discovered by other means, but also by providing new leads of its own. Finally, as any method of data management and analysis becomes accepted, it tends to be reflected in more standardized and efficient methods of data collection, with the result that a greater percentage of the work reported is then in a form suitable for inclusion in the system.

FILE CONTENTS AND STRUCTURE

Expressed in general terms, eq 1 and 2 are deceptively simple. However, it should be noted that three quite different types of information need to be expressed: descriptive, structural (topological), and numeric. As will be seen in later examples, the system anticipates the need to use Boolean logic dealing in all three types of information. The basic structure of the most fundamental unit in the file, the "Set," is shown in Figure 1. A "Set" consists of all the information associated with a given biosystem for which one or two regression equations suffice to rationalize the variation in activity observed for the series of compounds reported. A "Set" may contain information from several sources as long as the experimental conditions are comparable. Within each "Set" are found three main subclasses of information:** descriptor data, compound data, and regression analysis data. The computer arranges the analyzed information in the form shown in Figure 2.

DESCRIPTOR DATA

1. The "system" is the name of the biological entity being acted upon. Whenever specified, the various levels of complexity are stored in ascending order; *e.g.*, mitochondria liver mouse or hydrolase pepsin porcine.‡ All organisms are entered in the singular form; *e.g.*, "mouse" not "mice," and entries are made after referring to a Thesaurus. For example, the Thesaurus indicates that "bovine" is the preferred term for cow, cattle, ox, calf, bull, etc. It is obvious that an on-line terminal with suitable "prompts" could expedite the entry of this information in a standardized form more readily searchable by a "word" or "word fragment."[9] At present, we take the extra effort to specify in a search *E* followed by *coli* to be assured that both *Escherichia coli* and *E. coli* would be located.

As experience is gained in addressing practical questions to the file, a subdivision of the "system" field appears to offer some advantages. A simple subdivision being tested at present assigns dual alphanumeric characers as follows:

01 Nonenzymatic macromolecules; *e.g.*, albumin, fibrinogen
02 Enzymes
  A = Oxidoreductases
  B = Transferases
  C = Hydrolases
  D = Lyases
  E = Isomerases
03 Organelles; mitochondria, chloroplasts, etc.
04 Single cell organisms
  A = Algae
  B = Bacteria
  C = Cells (in culture)
  E = Erythrocytes
  F = Fungi (molds)
  P = Protozoa
  T = Tumor cells
  V = Virus particles
  Y = Yeasts
05 Isolated parts of organs; lobster axon, frog heart
06 Large functioning organism
  A = Animal
  H = Human
  B = Plant
  S = "Sectionized" animal; thyroidectomized dog

Example: Bacteria *in vitro* = 04B
  Bacteria *in vivo* = 04B, 06A

Obviously, the segregation stresses user interests rather than strict taxonomy.

2. The *compound*, that is, the name of "drug-X" in eq 1, can be entered either as text (*e.g.*, para-substituted benzyl alcohols) or as typed structure (*e.g.*, *p*-X-$C_6H_5CH_2OH$). The form is not critical but should be easily understood by the chemist studying the printed output. At the present time the logical OR of the WLN screens[10,11] from the "Compound Data" level is saved at this level as a "superscreen" and materially reduces the time needed for machine searching for specific molecular structure or fragments.

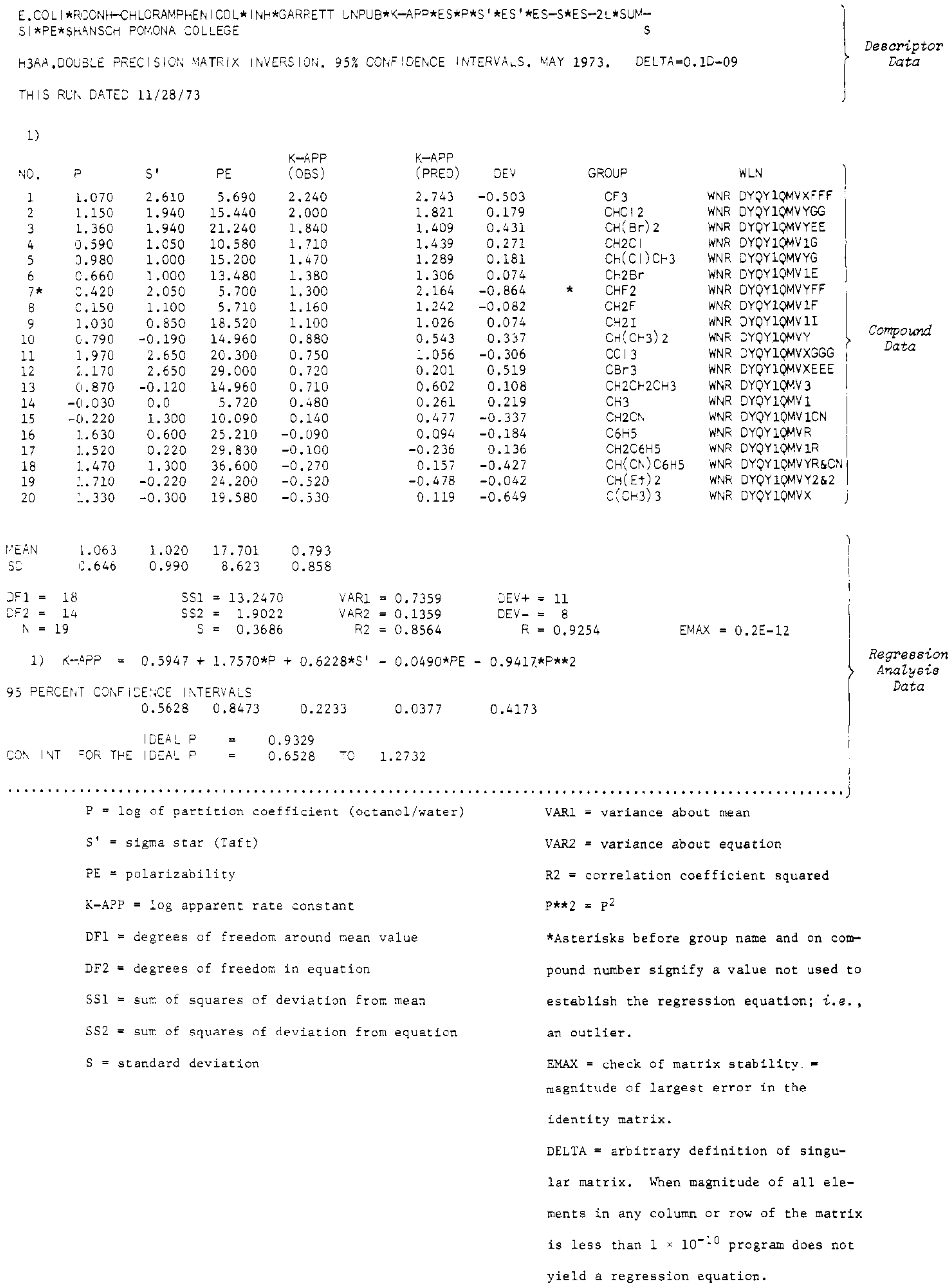
3. The *action* entered should provide sufficient understanding of the experimental measurement so that referral to the original publication is rarely necessary. Conditions of pH, temperature, and enzyme concentration should be stated whenever they are critical. However, until a greater standardization is made among biochemists and pharmacologists reporting their measurements, the usefulness of searching this subfile will definitely be limited unless carefully combined with the requirements made of other subfiles. For instance, "I 50," in reference to a compound acting against fungi, can mean that the normal growth rate was reduced to one-half, or it can mean that only 50% of spores survived and were viable. The same I 50 in an enzyme experiment can mean a 50% reduction in reaction rate at a given substrate concentration.

When the action is expressed as a variable response to a constant dosage (*e.g.*, per cent survival with dosage of 1 mg/kg), the data are of limited value for regression analysis. Cluster analysis and discriminant analysis might be better suited in dealing with data of this type.

4. In the *reference* field are entered the journal, volume, page, and date for published data. It may seem trivial, but it is nonetheless very expedient to standardize the abbreviation, spacing, and punctuation of this entry because only

** The mathematical operations attempted, whether or not they produced an acceptable equation, constitute a fourth subclass of information which may be deemed worthy of search in the future.

‡ Logically, it might be considered proper to extend the "system" in enzyme reactions to cover the substrate type and concentration as well as the conditions of temperature and pH. In practice, it appears more practical to include them as part of the "action."



```
E.COLI*RCONH-CHLORAMPHENICOL*INH*GARRETT UNPUB*K-APP*ES*P*S'*ES'*ES-S*ES-2L*SUM-
SI*PE*$HANSCH POMONA COLLEGE                                              S

H3AA.DOUBLE PRECISION MATRIX INVERSION. 95% CONFIDENCE INTERVALS. MAY 1973.   DELTA=0.1D-09

THIS RUN DATED 11/28/73

 1)
                                  K-APP              K-APP
NO.    P        S'       PE       (OBS)              (PRED)    DEV          GROUP          WLN

 1    1.070    2.610    5.690    2.240               2.743   -0.503         CF3          WNR DYQY1QMVXFFF
 2    1.150    1.940   15.440    2.000               1.821    0.179         CHCl2        WNR DYQY1QMVYGG
 3    1.360    1.940   21.240    1.840               1.409    0.431         CH(Br)2      WNR DYQY1QMVYEE
 4    0.590    1.050   10.580    1.710               1.439    0.271         CH2Cl        WNR DYQY1QMV1G
 5    0.980    1.000   15.200    1.470               1.289    0.181         CH(Cl)CH3    WNR DYQY1QMVYG
 6    0.660    1.000   13.480    1.380               1.306    0.074         CH2Br        WNR DYQY1QMV1E
 7*   0.420    2.050    5.700    1.300               2.164   -0.864    *    CHF2         WNR DYQY1QMVYFF
 8    0.150    1.100    5.710    1.160               1.242   -0.082         CH2F         WNR DYQY1QMV1F
 9    1.030    0.850   18.520    1.100               1.026    0.074         CH2I         WNR DYQY1QMV1I
10    0.790   -0.190   14.960    0.880               0.543    0.337         CH(CH3)2     WNR DYQY1QMVY
11    1.970    2.650   20.300    0.750               1.056   -0.306         CCl3         WNR DYQY1QMVXGGG
12    2.170    2.650   29.000    0.720               0.201    0.519         CBr3         WNR DYQY1QMVXEEE
13    0.870   -0.120   14.960    0.710               0.602    0.108         CH2CH2CH3    WNR DYQY1QMV3
14   -0.030    0.0      5.720    0.480               0.261    0.219         CH3          WNR DYQY1QMV1
15   -0.220    1.300   10.090    0.140               0.477   -0.337         CH2CN        WNR DYQY1QMV1CN
16    1.630    0.600   25.210   -0.090               0.094   -0.184         C6H5         WNR DYQY1QMVR
17    1.520    0.220   29.830   -0.100              -0.236    0.136         CH2C6H5      WNR DYQY1QMV1R
18    1.470    1.300   36.600   -0.270               0.157   -0.427         CH(CN)C6H5   WNR DYQY1QMVYR&CN
19    1.710   -0.220   24.200   -0.520              -0.478   -0.042         CH(Et)2      WNR DYQY1QMVY2&2
20    1.330   -0.300   19.580   -0.530               0.119   -0.649         C(CH3)3      WNR DYQY1QMVX

MEAN      1.063    1.020   17.701    0.793
SD        0.646    0.990    8.623    0.858

DF1 =  18              SS1 = 13.2470      VAR1 = 0.7359       DEV+ = 11
DF2 =  14              SS2 =  1.9022      VAR2 = 0.1359       DEV- =  8
  N = 19                 S =  0.3686        R2 = 0.8564          R = 0.9254          EMAX = 0.2E-12

  1)  K-APP  =  0.5947 + 1.7570*P + 0.6228*S' - 0.0490*PE - 0.9417*P**2

95 PERCENT CONFIDENCE INTERVALS
                 0.5628   0.8473    0.2233     0.0377     0.4173

                 IDEAL P     =     0.9329
CON INT  FOR THE IDEAL P     =     0.6528   TO   1.2732
.............................................................................................
```

P = log of partition coefficient (octanol/water)

S' = sigma star (Taft)

PE = polarizability

K-APP = log apparent rate constant

DF1 = degrees of freedom around mean value

DF2 = degrees of freedom in equation

SS1 = sum of squares of deviation from mean

SS2 = sum of squares of deviation from equation

S = standard deviation

VAR1 = variance about mean

VAR2 = variance about equation

R2 = correlation coefficient squared

P**2 = P^2

*Asterisks before group name and on compound number signify a value not used to establish the regression equation; *i.e.*, an outlier.

EMAX = check of matrix stability = magnitude of largest error in the identity matrix.

DELTA = arbitrary definition of singular matrix. When magnitude of all elements in any column or row of the matrix is less than 1×10^{-10} program does not yield a regression equation.

Figure 2. Computer report.

then does a simple alphanumeric sort make it possible to quickly ascertain if a particular set of data has been studied before.

5. The *dependent variable,* unless otherwise noted, is always given in log terms and, whenever possible, as $1/C$ where C is concentration in moles/liter. It is sufficiently precise to consider moles/kilogram as equivalent to moles/liter. When test results are given on a relative basis, RBR (log relative biological response) is the descriptor. In enzymatic studies, $1/K_m$ is used as the log of the Michaelis constant. In binding studies the log (% bound/% free) is reported as B/F.[12]

The *independent variables* are a set of parameters which attempt to describe the variation in the molecules causing the observed perturbation. Their names and derivations can be found in the references cited earlier.[2–7]

COMPOUND DATA

1. The numerical values of the activities are nearly always entered in log terms. Concentrations are converted to reciprocals so that the higher numbers denote higher activity.

2. Parameter values for the *entire molecule* can be entered in the case of log P, but more frequently they are entered as the value for the variable substituent group in a congeneric set. They are obtained from literature sources.[13] In principle, the extrathermodynamic parameters for a given substituent are invariant; for example, for $-CF_3$, $\pi = 0.88$, $\sigma_m = 0.43$, $\sigma_p = 0.54$, $E_s = -1.16$, etc. Thus it would seem desirable to have the machine load in, from disk storage, the selected parameter values as soon as the appropriate substituent structures are specified in WLN. At present this is not completely satisfactory, even for the physicoorganic systems in the file. The biological systems usually involve more complex substituents for which values of every type are not available and often do not correspond as closely to the model systems from which the parameters were derived as one might wish. Human selectivity is deemed essential at this stage.

3. The entire structure of "drug-X" is entered in WLN. Machine loading of parameters requires that the WLN for the substituent also be entered. As long as the entire file is not an order of magnitude greater than at present (*ca.* 2500 sets), the file structure shown in Figure 1, which requires that in a search for a given molecular structure each set must be examined at the "Descriptor Data" level, is not overly wasteful of time. Direct access to the compound data file with sorted WLN's and proper cross-referencing is planned for the future.

REGRESSION ANALYSIS DATA

The contents of this subfile are developed by the regression program employing double precision matrix inversion. A typical example is seen in the report shown in Figure 2. The types of data saved for possible searching are:

(1) number and types of parameters used in "best" equation
(2) operations used
(3) intercept
(4) coefficients of linear and power terms
(5) ideal P value
(6) correlation coefficient
(7) standard deviation
(8) confidence intervals
(9) number of compounds in regression

The selection of the "best" equation, like the loading of parameter values, *could* be made automatic, but it is not recommended. A program feature frequently used when the number of parameters is less than 12 generates equations for all possible combinations of terms, printing out only a summary of the equations and related statistics. Those with highest R and lowest S values are further examined to see that the terms are significant at >95% confidence level. These are then printed out in full, and the deviations of each drug in the set are examined.

It is often necessary to delete one or more data points of a set of congeners in a regression study. So that these data will not be overlooked and so that comparison of their fit with other points can be made, an option is provided for marking data cards with an asterisk. The "deleted" data are not used in deriving the equation, but the deviation from predicted activity is printed out for comparison (see CHF2 in Figure 2). By this equation selection procedure it is usually possible to settle on just one equation having at least four (preferably five) data points per term.

Before deciding upon the true significance of any parameter, it is prudent to check its degree of collinearity with the others. For each set studied, the program displays the correlation matrix for all the parameters and also calculates the T matrix.[14,15]

USE OF SEARCH-MANAGEMENT SYSTEM

Some of the utility of the present system for structure–activity studies can be illustrated by the following questions which, for ease of understanding, have not been phrased in the PL/I command language.

I. Query: What is the range of ideal log P's of phenolic bactericides?

Procedure: Examine subfile "System-4B" of "Descriptor Data" having Compound superscreen unspaced characters = "*Q*" and ["*R*" or "*L&J*" or "*T&J*"] and "Equation Data" with "Ideal *P*."

Print: Systems and Ideal P ordered on Ideal P.

II. Query: (a) What systems are strongly perturbed by the $-SCONH_2$ function:

(b) Which compounds containing this group show a high activity even when nonspecific hydrophobic interactions are unimportant?

Procedure: (a) Consider only those systems where the activity is reported as log $1/C$ and is greater than 4.0; therefore, "Descriptor Data," compound superscreens: uc (unspaced character) = S, V, and Z; dependent variable field = $1/C$. "Compound Data," WLN string = "SVZ" or "ZVS"; $1/C \geq 4.0$.

Print: "Systems" in alphabetic order.

(b) For each compound "hit",in (a) above, if P or PI does not appear in the equation or if $P < 2.0$.

Print: Alphanumeric sort of WLNs followed by "P" value and "System."

III. Query: (a) In considering perturbations due primarily to nonspecific hydrophobic interaction, is there any "natural" grouping based upon the systems' sensitivity to hydrophobic effects (as compared to the octanol/water model system)?

(b) What is the intrinsic activity scale of various pharmacophoric groups when any change in the remainder of the molecule has only a hydrophobic bonding effect?

Procedure: (a) "Equations" searched for those whose sole independent variable is PI or P.

Print: In order of increasing coefficient of independent variable. This list can be scanned manually for areas of highest "interval population," or else a "cluster analysis" procedure can be applied to sort it in three or more groups.

(b) As above, but eliminate sets having "MISC" in the compound field of "Descriptor Data." (For this type of question it would be advantageous to have entered in the compound field the WLN for the parent [unchanging] structure of the set.)

Print: In decreasing value of "intercept" the intercept, the parent WLN, and the "system."

IV. Query: Are there any substructures which the presently available parameters characterize only poorly; *i.e.*, are there any frequently encountered substructures in compounds which deviate greatly from predicted activity?

Procedure: The standard deviation value "S" in the "Equations" section of each set is compared with the "Dev" value for each of the compounds in the set; WLN saved for every compound where "Dev" $\geq 3\ S$.

Print: A permuted list of the WLN's, the number found in each permutation type, and the per cent of the total list represented by that permutation. This is compared to a permuted list of the entire compound file which also lists per cent of each permutation. Permuted groups (substructures) occurring with twice the frequency in the deviant list as in the entire list should be considered as inadequately characterized.

ACKNOWLEDGMENT

This work was supported by Grant CA 11110 from the National Institutes of Health Contract NIH DRR-70-4115 from the National Institutes of Health and the Pomona College Medicinal Chemistry Project.

LITERATURE CITED

(1) Ariëns, E. J., "A General Introduction to the Field of Drug Design" in "Drug Design," Vol. I, E. J. Ariëns, Ed., Academic Press, New York,

N. Y., 1971, pp 2–263.

(2) Hansch, C., "Quantitative Approaches to Pharmacological Structure-Activity Relationships" in "Structure-Activity Relationships," Vol. I, C. J. Cavallito, Ed., Pergamon Press, Oxford, England, 1973, pp 75–165.

(3) Bruice, T. C., Kharasch, N., and Winzler, R. J., "A Correlation of Thyroxine-Like Activity and Chemical Structure," *Arch. Biochem. Biophys.*, **62**, 305–317 (1936).

(4) Free, S. M., Jr., and Wilson, J. W., "A Mathematical Contribution to Structure-Activity Studies," *J. Med. Chem.*, **7**, 395–399 (1970).

(5) Craig, P. N., "Comparison of the Hansch and Free-Wilson Approaches to Structure-Activity Correlation," *Advan. Chem. Ser.*, No. 114, 115–129 (1972).

(6) Verloop, A., "The Use of Linear Free Energy Parameters and Other Experimental Constants in Structure-Activity Studies" in "Drug Design," Vol. III, E. J. Ariëns, Ed., Academic Press, New York, N. Y., 1972, pp 133–187.

(7) Kier, L. B., "Molecular Orbital Theory in Drug Research," Academic Press, New York, N. Y., 1971.

(8) Reference 1, p 24.

(9) Colombo, D. S., and Rush, J. E., "Use of Word Fragments in Computer-Based Retrieval Systems," *J. Chem. Doc.*, **9**, 47–50 (1969).

(10) Smith, E. G., "The Wiswesser Line-Formula Chemical Notation," McGraw-Hill, New York, N. Y., 1968.

(11) Granito, C. E., Becker, G. T., Roberts, S., Wiswesser, W. J., and Windlinx, K. J., "Computer-Generated Substructure Codes (Bit Screens)," *J. Chem. Doc.*, **11**, 106–110 (1971).

(12) Bird, A. E., and Marshall, A. C., "Correlation of Serum Binding of Penicillins with Partition Coefficients," *Biochem. Pharmacol.*, **16**, 2275–2290 (1967).

(13) A computer-sorted listing, updated semiannually, of log *P* values and electronic and steric parameters for over 800 substituents is available. Inquiries may be addressed to: A. Leo, Pomona College Medicinal Chemistry Project, Department of Chemistry, Pomona College, Claremont, Calif. 91711.

(14) Farrar, D. E., and Glauber, R. R., "Multicollinearity in Regression Analysis: The Problem Revisited," *Rev. Econom. Statistics*, **49**, 92–107 (1967).

(15) Haitovsky, Y., "Multicollinearity in Regression Analysis: Comment," *Rev. Econom. Statistics*, **51**, 486–489 (1969).

Computerized Management of Structure–Activity Data. II. Decoding and Searching Branching Chains and Multiplied Groups Coded in WLN

A. LEO,* DAVID ELKINS,† and CORWIN HANSCH

Department of Chemistry, Pomona College, Claremont, California 91711

Received December 21, 1973

As each WLN symbol for a structure containing a branching chain and/or multiplied groups is extracted in a left-to-right scan, the symbol to which it was connected in the *graphic* formula[1] must be known. For highly branched structures, especially where ring systems and/or multipliers are present, the program logic becomes quite complex. Ring substituents are discussed but decoding of the ring structure itself is reserved for the following article.

The increased use of Wiswesser Line Notation (WLN)[2] for the computerized storage, sorting, and searching of chemical structures[3-5] has created the need for programs that will either generate the notation from electronically drawn diagrams[6,7] or check the accuracy of manual encoding.[8] In the latter case the algorithms needed to properly dissect the WLN and compare it to a molformula are also useful to provide information which makes subsequent sorting and structure searching more efficient. The following WLN decoding programs are possibly similar to those in use elsewhere, but the particular algorithms for handling branched chains, multiplied groups, and rings have not been widely publicized, and they may prove helpful to those contemplating their own computerized structure files.

DISSECTION

The first step in the decoding process is the resolution of the various components of the WLN. The machine must be programmed to extract locants, multipliers, ring kernels (the kernel is that portion of a ring system set off by the ring initiator, D, L, or T, and the closing J), alkyl chain numerals, two-letter atom symbols (–NA–, –FE–), etc., from the string of simple structural symbols which comprises the essence of the WLN. This is accomplished by examining each character in its context during a left-to-right scan of the notation. When locants or ring systems are encountered, special routines are called upon to extract these items.

ATOM EQUIVALENCE

The next step is to convert the extracted WLN symbol into its atomic equivalent. Some WLN and atomic symbols are identical, such as I, O, and S, while others, such as M, V, and Z, denote multiatom fragments. Except for ring kernels which are handled by special routines (see the following paper in this series), conversion is made *via* a Symbol Equivalence, Table I.

Note that methyl-contracting branches (K, X, and Y), when they are not within rings, are entered with their full complement of methyl groups. Ring positions which are not specified as locations for heteroatoms or V, X, or Y are considered initially to be fully substituted with hydrogen atoms. This includes the benzene symbol R. An alkyl chain numeral, *n*, is entered as C_nH_{2n+2}.

VALENCE AND BOND FORMATION

The valence of each symbol (*e.g.*, the number of single chemical bonds available for attachment to other symbols) is determined from Table II.

* To whom correspondence should be addressed.

† Present address: G. D. Searle & Co., Chicago, Ill. 60680.