Computer-Managed Automatic Data Retrieval and Prognosis System for Rate and Equilibrium Constants of Organic Reactions

V. PALM

Department of Organic Chemistry, Tartu University, Küütri 20, Tartu, 202400, Estonia, USSR

Received August 10, 1990

A system suited for the automatic retrieval of rate and equilibrium constants and related data from databanks or for the calculation of these values by using parametrized correlation equations is described. The search and identification of structures and reactions is based on the specific language LINCS for linear coding of structures via their representation as labeled graphs. A specially suited multiparameter combined linear and nonlinear least-squares data-processing procedure for the parametrization of models is included.

1. THE GENERAL CONCEPT

Any science has to meet some basic requirements and restrictions. Otherwise, it meets the hazard of having little to do with reality. Put simply, it is necessary to achieve a reproducible and adequate empirical and theoretical description of some observed space. This space is defined as a dependence of some measurable quantity on a definite combination of levels of significant physical factors (e.g., temperature, pressure, solvent, substituents, etc.). These constitute a set of the coordinate axes of that space. Both the measurable quantity and the position of any of the axes are characterized by some logical, discrete (integer), or real value, and a set of these for all axes specifies a point in the observed space. Such representation of this space is obligatory and sufficient for any empirical approach, but not always for a theoretical one. In general, the theoretical conception will define how some not directly usable logical or discrete factor levels (e.g., the formula or sequence number of the substituent or solvent) have to be transformed into meaningful quantities (i.e., definite values of some arguments whose scales are defined in the framework of the conception). On the other hand, the primary representation of the factor levels and the measurable quantity are meaningful in the framework of some other theoretical concepts—let us call them "observation models" to distinguish them from "interpretive" ones devoted to the theoretical description of the observed space (e.g., for predictive purposes).

Interpretive models usually contain a set of parameters whose values may be estimated only by using the empirical data for the target observation space.

The applicational aspect of the science may be reduced to the description of observation spaces from which constituents are needed for a design of some kind of technology. For multidimensional observation spaces, the total number of distinguishable points becomes practically infinite, and the real importance of interpretive models is highly increased. Therefore, simple systems for automatic retrieval from a corresponding database become quite ineffective for such cases because of too frequent negative search results.

In the field of chemistry all structure-dependent data are defined via multidimensional observation spaces. This is especially true for reactivity and analogous data due to significant additional interference of the temperature, medium, and other factors. Most strictly, the reactivity data are represented by rate and equilibrium constants. Therefore, in the mid-1960s, this project[1] was initiated to design and realize a combined computer-managed system for search and retrieval, checking and parametrization of the interpretative model usable for prognosis, proceeding from the database for rate and equilibrium constants of chemical reactions.

2. THE DATABASE

2.1. Tables. As a base for the whole project, tables of compiled data are used.[2-16] The preparation of complementary issues is ongoing.

These rate and equilibrium data cover several reaction types as follows: dissociation of the hydrogen acids according to the Brønsted scheme, rate constants for proton transfer, first-order nucleophilic substitution at nonaromatic centers and solvolysis, second-order nucleophilic substitution at nonaromatic centers, electrophilic substitution at nonaromatic centers, addition to double and triple bonds, elimination with the formation of double and triple bonds, hydrolysis of carboxylic esters, reactions of carbonyl compounds as electrophiles, electrophilic aromatic substitution, aromatic nucleophilic substitution, tautomeric equilibria, intramolecular rearrangements, formation of complexes from general acids and bases, substitution of ligands for compounds of the elements from rows II and III, and complexing of different elements with ligands.

Altogether, more than 300 000 independent values of constants have been compiled. Data are structuralized according to a hierarchy of factors as follows: reaction type, definite reaction solvent, substituents at the reaction center, temperature, and specific conditions (e.g., nature and concentration of species added).

2.2. Computer-Manageable Databank. For the input, analysis, and retrieval of data, the specific language LINCS for linear coding of structures and reactions has been developed.[1,17-19] Structures are represented via labeled graphs. The nodes of these correspond to single atoms, and the arcs represent covalent bonds including the dummy ones. Actually, nodes may represent some structural unit of arbitrary complexity if the corresponding definition is present in a table of "direct codes". Reactions are specified by the use of specific labels for the bonds to be formed or broken. These reflect the nature (hetero- or homolytic) and sign of the corresponding electron pair shift or heterolytic processes. For condensed cyclic systems the method proposed by Vleduts and Geivandof[20] is employed.[21]

The possibility of unlimited introduction of new direct codes opens a convenient way to adjust the input language as favored by the user.

LINCS codes are transformed by the program into full graph tables, which are used internally for the storage and processing of the information dealing with reactions and structures.

The commands for searching the reactions and related information also have to be formulated via LINCS codes. A single definite reaction or a set (reaction series) with arbitrary variable factors may be searched. The whole information stored for the row of the data table corresponding to a given value of the rate or equilibrium constant becomes available for any further use. It could be a simple compilation and bibliography in the scope of information ordered. But it is also possible to order the formation of initial data sets for further

0095-2338/90/1630-0409$02.50/0

(statistical) treatment by making use of another procedure.

3. PROCEDURE FOR CHECKING AND PARAMETRIZATION OF MODELS

This procedure has not yet been fully described, and only certain features have been cited in publications (e.g., refs 22–24) which deal mainly with the results of its use for data processing. It may be used either completely independently or in combination with the search system and databank.

Different modes for representation, preparation, and storage of initial data are available for use. The most general representation of the relationship between some measurable (dependent) quantity and the levels of the set of factors influencing it could be employed. Using this, it is possible to compile completely independent databases for this procedure. For this version, the real values of arguments are stored in a special (stationary) bank representing a matrix. The columns of the last one correspond to different argument scales and are related to different potentially significant factors. The row indices represent sequence numbers of the levels of corresponding factors.

The additional argument scales formed as products (cross terms) may be declared via corresponding citation of the already defined ones.

The data sets represent experimental values to be processed and the sequence numbers for the levels of factors, the significance of which has to be tested. The flexible mode to define the correspondence between data set factors, sequence numbers, and the stationary bank columns enables one to use such a bank for unlimited numbers of several data sets and to define several models and select different subsets of data rows for each set.

Models are checked, and the parameters, standard deviations (SD), and other statistics are calculated in the framework of multiparameter linear or nonlinear least squares[25-28] (LLSQ and NLSQ) combined into a single procedure.

3.1. LLSQ Procedure. The standard principal approach via formation of the correlation or normalized covariation (if the intercept is absent) matrix, its rearrangement,[29] and the calculation of the results and statistics is normalized (in the units of dispersion square roots) and natural scalings are used. For a single job, the matrix is formed only once. All the operations in the course of calculation of the value of its determinant and rearrangement touch only the elements of the upper triangle and the diagonal ones. The exclusion or restoration of data rows is realized via matrix recalculation, and the exclusion of argument column is realized via equalizing to zero the corresponding nondiagonal column and row elements for the upper triangle. The initial state of the upper triangle of the matrix is easily restorable because its lower triangle is preserved unchanged.

The main essential problems are the specification of the model by the choice of the set of statistically significant argument scales and the extraction from the initial data set of a statistically self-consistent subset purged from significantly deviating rows. The possibility exists to order the sequence of the execution of those procedures. For both of them different confidence (or risk) levels may be declared, and the exclusion of significantly deviating rows is performed on different, gradually enhancing, risk levels.

Because of the statistical significance, the argument scales should not be excessively linearly interrelated. The last requirement excludes the use of stepwise regression as the scale most dependent on several significant ones which may happen to be selected prior to the last ones even if its own real contribution is less or even absent. Therefore, the principle of the reduction of the possible abundant starting model is employed in the limits allowed by the requirement to grant the presence of the statistical degrees of freedom. The significance of the coefficient with maximum relative standard deviation (SD) is tested by using the Fisher criterion. The corresponding argument scale is excluded from the model if this test confirms that it is nonsignificant on the given risk level.

The interrelation of argument scales is caused by their nonorthogonality and leads to the increase of the relative uncertainty of coefficients at the expense of each other (the "overpumping" effect; OVEF), which is not reflected in the SD value for the description of data. At extreme cases the loss of precision in the course of rearrangement of degenerated matrices are reflected by low or zero values of its determinant (D). On the other hand, the lower the D value for the extended (via inclusion of the dependent quantities in the set of argument scales) matrix ($D0$ value), the better the description properly reflected by the ratio $RD = D0/D$.

The extent of the increase of RD value as a result of the exclusion of a definite argument scale serves as a criterion of its statistical significance and is used for the preliminary reduction of the model before the regression procedure is started. Special procedures are included to exclude nonsignificant argument scales from the model if $D0$ (or both $D0$ and D) are indistinguishable from zero at a given precision loss tolerance.

If the solution is obtained, the total OVEF is reflected by the value of the ratio $TR = \sum S \times O(J)^2/SO^2$ equal to the trace of the rearranged correlation (covariation) matrix and the partial one for a definite scale—by ratio $OP(J) = S \times O^2/SO^2$. $S \times (J)$ and SO denote the normed values for the SD of a Jth coefficient and a total one. The OVEF is remarkable if $TR > 1$, and the statistical significance of the scale with the maximum SO value is subjected to the Fisher test if its $OP(J) > 1/NP$ where NP denotes the total number of parameters to be estimated. If it appears to be significant, but the relation $SO(J)/SO < \sqrt{1 + (TR - 1)STD}$ is satisfied, the exclusion of this scale is confirmed due to its significant partial OVEF. $SO(J)$ and STD denote the SO value after exclusion on the Jth scale and Student's criterion value, respectively.

Significantly deviating rows are excluded according to the Student's test by using the statistics obtained after exclusion of the row under consideration, if its absolute value of deviation exceeds the criterion defined as a maximum range for possible experimental error.

The formation of cross terms from natural basic argument scales induces additional OVEF the higher and larger the distances between the centers and the origins of the scales are. Therefore, the possibility exists to define the cross terms as formed from centered basic component scales. The final result is also calculated for original scales.

3.2. NLSQ Procedure. Nonlinear models are represented by corresponding blocks or subprocedures. Several of them may be compiled for use; each one being identified by a sequence number. Several standard models are included in the system from the very beginning. Some of these are cited below.

One or more argument values represented by the definite position in a stationary bank can be declared as additional parameters to be estimated. If this is done for a model initially formulated as linear, the automatic transformation to the corresponding nonlinear model with sequence number zero occurs (no special nonlinear subprocedure is needed).

A model is available for checking the hypothesis that the experimental rate constants for a given series are actually the sums of terms which correspond to parallel paths of the process (e.g., Sn1 and Sn2 mechanisms for nucleophilic substitution). It is represented by the general relationship:[30]

$$\ln K = \ln \left(\sum_K C(K) \exp F(K)\right)$$

where $C(K)$ denotes concentrations of some ingredients or is

equal to one, $F(K)$'s are (multi)linear functions and represent the dependence of the logarithm of the rate constant for the Kth path on a set of arguments (substituent or solvent constants, temperature, etc. and corresponding cross terms). The intercept and coefficients of these functions form a set of parameters to be estimated.

The procedure for the digital solution of formal kinetic problems formulated for a certain mechanistic scheme includes in the general case fast equilibria and irreversible and reversible slow stages. The parameters for estimation are rate and equilibrium constants.

For every subprocedure, the possibility is reserved to compile a specific preliminary procedure, e.g., for the calculation of the initial approximation for the set of parameters to be estimated.

The NLSQ procedure itself is realized as the local minimization of the SD value in the restricted solution space formed by coordinates related to the parameters for estimation. Two kinds of limits are introduced—the "mathematical" and the "physical". First the system coordinates are protected from obtaining values which lead to mathematical error interruption of calculations. Then limits are introduced to avoid meaningless or "impossible" values. The mathematical limits are protected from being violated anyway. For physical limits, the version may be chosen when the passage of limits in the course of the SD minimization is only checked and registered. If at the minimum point some limits remain violated, the limit value for a single coordinate is assigned by using the criterion of the minimum rise of SD. Violation of physical limits can be accompanied by the use of the punishment function.

For minimization, the parallel tangents method[31] is used. The starting direction for movement in a running minimization cycle may be defined by the use of either the solution of the linear problem proceeding from the Jacoby matrix formed by the digital method of calculation of derivatives or via straightforward detection of the antigradient direction for SD. The automatic change for a special version of stepping along a bottom of a deep "valley" is available to be ordered.

Special criteria for detection and significance of the OVEF are introduced for NLSQ. The squares of ratios $SXT(J)/SXS(J)$ of the SD of the Jacoby matrix coefficients calculated for the solution point by the use of multilinear $[S \times T(J)]$ and nonlinear $[S \times S(J)]$ in respect to the Jth coefficient are used. The sum of these ratios is used to characterize the total OVEF. Those parameters causing significant OVEF are removed from the model in addition to those which are not statistically significant according to the Fisher test.

A generally usable procedure for the calculation of initial approximations near the global minimum is available. This scans the set of points located on the definite surface surrounding the central one. From each such point the SD minimum is reached. From those whose locations do not coincide with the central point, the lowest SD is considered and, if it corresponds to a deeper minimum in comparison with the one for the center point, it is considered as a new center point. The procedure is finished when no such deeper minimum of SD is found. The last center point defines, therefore, the initial approximation at the presumed global minimum.

Combined linear–nonlinear models may be formulated. The intercept and coefficients of the linear part for any given point of the solution room formed by proceeding from the parameter of the nonlinear part of the function are only calculated by the LLSQ procedure.

Both LLSQ and NLSQ are able to solve problems proceeding from a single formulation of the abundant starting model only if enough data are available in the set treated to grant the presence of the statistical degrees of freedom (SDF). Otherwise the results for the set of alternative models have to be compared. If the situation of the lack of SDF should appear, the corresponding number of argument scales with higher sequence numbers is excluded from the formula of the linear model. In the case of the NLSQ problem, the parameters of the nonlinear part are excluded from the set ones for estimation and the constant values equal to the initial approximation are preserved.

4. PROCEDURE FOR THE PROGNOSIS OF CONSTANT VALUES

For every running state of the databank, the stationary bank of potential argument scales, and the list of models parametrized by the use of the parametrization procedure described, a list of models with definite values of parameters may be obtained. For the realization of the corresponding prognosis capacity, a package of procedures is available. This is suited for the storage and automatic retrieval of equations as well as the values of corresponding parameters and arguments and executing the calculations according to the models represented by the equations. This package is described elsewhere in greater detail.[32-35]

The modeling relationships employed up to the present are (multi)linear correlation equations for log K and are pK derivable by using the general formal theoretical approach to the quantitative interpretation of experimental data (see a more detailed account in ref 36 and literature cited therein). However, from the technical viewpoint, there are no restrictions on the use of any other types of models.

For trivial reasons, the predicting power of this program exceeds the informational capacity of the bank of experimental values by many orders of magnitude.

5. TECHNICAL IMPLEMENTATION

The programs are written in FORTRAN ANSI 77 and realized in this version for the NORD 100 computer (Norsk Data Co., Norway). A version in Microsoft FORTRAN under MS DOS for use with a PC AT series is under development. This version is intended to be the one suitable for distribution. A 40mb hard disc and 640kb RAM disc are sufficient for the installation of the system.

For the parametrization package, the dynamic memory distribution is realized, and the additional use of a 380kb RAM disc is enabled.

6. POSSIBLE APPLICATIONS

The use of the system described for the retrieval or prediction of rate and equilibrium constants is trivial. Beside this, and independent of the actual contents of the database, it presents an example of artificial "intelligence", enabling one to derive from the empirical data compiled in the databank the maximum possible conceptual and predictive generating information for the circle of ideas represented by models available for data processing. On the other hand, the list of data lacking for the proof of models available could be specified.

REFERENCES AND NOTES

(1) Kiho, J. K.; Vleduts, G. E. Concerning Some Preliminary Problems in the Use of Digital Computer for the Application of the Method of the Correlation Equations in Organic Chemistry. *Reakts. Sposobn. Org. Soedin.* **1965**, *2* (2), 88–107, in Russian, English summary.
(2) *Tables of Rate and Equilibrium Constants of Heterolytic Organic Reactions*; Palm, V. A., Ed.; Publishing House of VINITI: Moscow, 1975; Vol. I (1).
(3) *Ibid.* 1975, Vol. I (2).
(4) *Ibid.* 1976, Vol. II (1).
(5) *Ibid.* 1977, Vol. II (2).
(6) *Ibid.* 1977, Vol. III (1).
(7) *Ibid.* 1977, Vol. III (2).
(8) *Ibid.* 1977, Vol. IV (1).

(9) *Ibid.* 1977, Vol. IV (2).
(10) *Ibid.* 1978, Vol. V (1).
(11) *Ibid.* 1979, Vol. V (2).
(12) *Tables of Rate and Equilibrium Constants of Heterolytic Organic Reactions*; Palm, V. A., Ed.; Publishing House of Tartu State University: Tartu, 1984; Suppl. Vol. 1, Issues 1–2.
(13) *Ibid.* 1985, Suppl. Vol. 1, Issues 3–5.
(14) *Ibid.* 1987, Suppl. Vol. 2, Issues 1–3.
(15) *Ibid.* 1989, Suppl. Vol. 3, Issues 1–3.
(16) *Ibid.* 1989, Suppl. Vol. 4, Issues 1–3.
(17) Kiho, J. K. LINCS system. *Reakts. Sposobn. Org. Soedin.* **1970**, *7* (1), 94–111, in Russian, English summary.
(18) Kiho, J. K. Input of Ographs into Computer. *T. BC TGY* **1976**, *35*, 18–36, in Russian.
(19) Kiho, J. K. Derivation of Canonic Numeration of the Nodes of Graphs. *Ibid.* **1976**, *37*, 44–49, in Russian.
(20) Vleduts, G. E; Geivandof, E. A. *Automatic Informational Systems for Chemistry*; Nauka: Moscow, 1974, in Russian.
(21) Jalas, A. N. Codification and Input of Mosaic Systems in System "Cis-Tartu". *Abstracts of the VIII All-Union Conference on Use of Computers in Spectroscopy and Chemical Research*; Siberian Branch of Academy of Sciences of U.S.S.R.: Novosibirsk, 1989; p 153, in Russian.
(22) Leinbock, R. A.; Palm, V. A. Nonlinear Parametrization of Equations for Atomic Spectral Terms. 1. Specification of the General Character of Dependence on Orbital–Orbital Shielding Constants. *Org. React. (N.Y. Engl. Transl.)* **1983**, *20* (3), 373–396.
(23) Palm, N.; Palm, V. Determination of the Kinetic–Equilibrium Parameters of the Reaction of Triphenylaluminum with Benzophenone. *Org. React. (N.Y. Engl. Transl.)* **1985**, *22* (2), 131–141.
(24) Nummert, V.; Eek, M.; Palm, V. Kinetic Study of Alkaline Hydrolysis of Substituted Phenyl Tosylates. XIV. Discussion of Results of Kinetic Measurements in 80% Aqueous Dimethylsulfoxide. *Org. React. (N.Y. Engl. Transl.)* **1985**, *22* (3), 263–294.
(25) Bennett, C. A.; Franklin, N. L. *Statistical Analysis in Chemistry and Chemical Industry*; John Wiley: New York, 1967.
(26) Kafarof, V. V. *Methods of Cybernetics in Chemistry and Chemical Technology*; Chimia: Moscow, 1976, in Russian.
(27) Lvovskij, E. N. *Statistical Methods for Constructing of Empirical Formulae*; Vyshaja Shkola: Moscow, 1982, in Russian.
(28) Laurent, P.-J. *Approximation and Optimization*; Mir: Moscow, 1975, in Russian.
(29) Ageev, M. L.; Alik, V. P.; Markof, J. I. *Library of Algorithms*. Sovetskoe Radio: Moscow, 1976; pp 33–36, in Russian.
(30) Palm, V. A. Nonlinear Model of Substituent Effects for Parallel Reactions. *Abstracts of the VIII All-Union Conference on Use of Computers in Spectroscopy and Chemical Research*; Siberian Branch of Academy of Sciences of U.S.S.R.: Novosibirsk, 1989, pp 301–302, in Russian.
(31) Dashevskij, V. G. *Conformations of Organic Molecules*; Chimia: Moscow, 1974; pp 133–134, in Russian.
(32) Jüriado, T. J.; Palm, V. A. Program Package for Computer Storage and Automatic Search of Correlation Equations and for Calculation of Rate and Equilibrium Constants. 1. Digital Coding System of Equations of Chemical Reactions. *Org. React. (N.Y. Engl. Transl.)* **1984**, *21* (3), 255–284.
(33) Jüriado, T. J. Program Package for Computer Storage and Automatic Search of Correlation Equations and for Calculation of Rate and Equilibrium Constants. 2. Algorithm for Search of Index of Correlation Equation From Identification Arrays on the Basis of Reaction and Substituent Codes. *Ibid.* **1984**, *21* (4), 375–387.
(34) Jüriado, T. J. Program Package for Computer Storage and Automatic Search of Correlation Equations and for Calculation of Rate and Equilibrium Constants. 3. Algorithm for Search of Solvent and Temperature. *Ibid.* **1984**, *21* (4), 388–404.
(35) Jüriado, T. J. Program Package for Computer Storage and Automatic Search of Correlation Equations and for Calculation of Rate and Equilibrium Constants. 4. Algorithm for Calculation of Rate and Equilibrium Constants on the Basis of Results of Reaction Set Search. Short Manual of Program Use. *Ibid.* **1984**, *21* (4), 405–417.
(36) Palm, V. A. Foundations of Quantitative Theory of Organic Reactions, 2nd ed.; Chimia: Leningrad, 1977, in Russian.