Proc. Estonian Acad. Sci. Chem., 2006, **55**, 1, 3–13

Computerized database and software for retrieval, processing, and prognosis of rate and equilibrium constants of chemical reactions

Viktor Palm*, Natalia Palm, and Tiina Tenno

Institute of Chemical Physics, University of Tartu, Jakobi 2, 51014 Tartu, Estonia

Received 20 September 2005

Abstract. A computerized database on rate and equilibrium constants of chemical reactions was created. Two versions of the database, one limited to the pKa data, which employs the graphical display of the structures, and the other that enables systematic withdrawal of data selected by several conditions over all 14 main reaction types, are represented. The algorithm for statistical treatment of large sets of data to create and check mathematical models for their chemical interpretation was modified and refined. The software has been used for the treatment of large sets of data on solvent and subtituent effects and can be employed for the prognosis of new data not yet studied experimentally.

Key words: coding of chemical reactions, coding of chemical structures, computerized database, rate constants, equilibrium constants, multilinear mathematical models, correlation equations, non-orthogonality.

INTRODUCTION

The general scheme of scientific investigation consists of three main components: the experimental (observational) part, the construction of a compilation of the experimental data available in the scientific literature, and the interpretation of different data sets selected from this compilation via some systematic way. The more experimental results are in principle available, the more labour consuming the extraction of the data sets to be used for the data processing related to some (theoretical) models becomes. The models with the proved degree of reliability can be employed for the prognosis of the new data

* Corresponding author, viktor.palm@ut.ee

for the situations not yet studied experimentally [1]. Such prognosis is an essence of the use of scientific results for different practical applications. The use of the computerized approach for the automation of the stages of the process described, enabling the generation of primary experimental data, is the topic of this communication. This is realized on the example of the rate and equilibrium constants for the heterolytic chemical reactions. However, the technique described and the related software can be applied in other fields as well.

The traditional way of the general representation of existing quantitative experimental data is reduced to the compilation of the related information into tables published as a series of books ("hard copies"). For the field stated above, it is a result of the efforts of the scientific team at the Department of Chemistry of the University of Tartu during 1970–1990. The set of 16 volumes (10 basic and 6 supplementary ones) of the *Tables of Rate and Equilibrium Constants of Heterolytic Organic Reactions* was created and published [2, 3]. The number of the independent data rows covered reached nearly 500 000.

DATABASE OF RATE AND EQUILIBRIUM CONSTANTS

The extraction of the systematically organized data compilations from the Tables mentioned is a labour consuming and wearisome business. Especially, if one takes into consideration that closely related data are located in the basic as well as in the supplementary volumes. This is inevitable when new data becoming available are involved.

Therefore, from the very beginning of the use of computers a project was started for the generation of software for the computerized systematic storage and retrieval of such information. However, one has to keep in mind that the computers available at the beginning of the 1970s were hardly applicable for convenient practical use for this purpose. Only the last generations of powerful PCs enable to obtain an acceptable solution to this problem for an ordinary user.

The system described was created by combining separate parts as reported below.

For the storing of the reactivity information, a special language, *LINCS* (Linear Coding of Structures), has been created [4, 5]. This language represents the chemical structures of reactants via connected labelled graphs with atoms as the nodes and chemical bonds as arcs. All additional information about reaction conditions (logarithms of the values of rate or equilibrium constants, solvents, temperature, pressure, chemical ingredients added, methods of measurement, and literature sources) is represented by additional nonconnected labelled nodes. The possibility of distinguishing of the reaction centre from the remaining part of the corresponding molecule (for instance, the substituent) is introduced. The information represented via *LINCS* codes is automatically converted into graph tables used for the internal storage.

Secondary coding via introduction of direct codes (*DC*) for arbitrary substituents (discrete parts of the molecules) is introduced. Users can voluntarily define these *DC*, representing any fragments of molecules by single nodes. They can be used for a more compact representation of the *LINCS* codes, employing the conventional symbols used by chemists. For each *DC* a sequence number can be introduced. This makes possible an alternative numerical coding of (structural) information, which is the preferred way of the representation of the information when the sets of initial data for (statistical) processing have to be created. For instance, the *DC* denoted as ET (or Et) can be substituted for the basic *LINCS* code (<<< C.H > .H >. <<< C.H > .H > .H >>) for the fragment (substituent) called ethyl (CH_3–CH_2–). No limit exists for the complexity of structural units for which the *DC* can be introduced but the maximum amount of the memory reserved for the storage of a single *LINCS* code.

For the retrieval of information from the database, two different versions of software are available. One of them represents a dialogue version of the selection of a branch of the tree of hierarchically organized reactant structures, with the possibility of defining the limiting restrictions for the selection of the solvent, temperature, and pressure. This version is realized [6] for the Brönsted acidity–basicity and equipped with the graphical representation of the chemical structures. Its use is reduced to the sequence of choices of the structural types and solvents until the terminating branch is reached. The results of each session of data retrieval are visualized graphically on the computer screen and saved for subsequent fast reproduction, making use of the computer program specially suited for this purpose. Additionally the results of each session are stored in a plain text file. This version enables visual observation of the whole procedure of the ordering of data and the results of their retrieval. Therefore, it is suitable for the purpose of training students. However, it is not very convenient for the creation of specifically organized data sets for their further use for some kind of processing (checking the degree of their conformity to some mathematical model etc.).

For the last purpose, the other version for the retrieval of data has been created. This covers all types of the reactions represented in the database. The more general types of reactions are identified with basic digital codes (sequence numbers *M*). Altogether 16 such general reaction types are covered by the database. More concrete subtypes of reactions are labelled by the secondary digital codes (*N*) and the complete reaction code is represented by their combinations like *M/N*, which corresponds to the *LINCS* code of some definite structure of the reaction centre and its change in consequence of the respective chemical process. Special labels are in use for the denotation of the breaking and forming of bonds and the way (polarity) of the bond breaking and formation. The set of digital codes is related to “elementary” substituents (any one can be declared as such if the corresponding *DC* is declared). For complex substituents, the bridging fragments or their sequences can be introduced. These bridges connect some elementary substituent with the reaction centre. They are defined via corresponding bivalent structural fragments represented by respective *DC* and related digital codes. For instance, *DC* as 1Ph4 for the *p*-phenylene bridge with *LINCS* code – <:C:CH::C:CH::>[4.1 – (for <C.H> the *DC* for CH is already substituted) can be used. The order for data retrieval consists of the reaction code *M/N* or reaction *LINCS* code with actual or variable substituents connected to the (indexed) position(s) at the reaction centre. Additionally, the list of sequence number(s) (index(es)) of the substituent(s), and of the (components of) solvent can be specified. Solvents can be represented by the corresponding *DC* and sequence number (solvent index), too. It is also possible to specify the value of molar percentage or other concentration measure stored for a particular data set for the second component of the binary mixed solvent, the temperature (range), the pressure (range), and other restrictions for the experimental method, and different notes and the literature sources for data rows to be retrieved. If these restrictions are not specified, all data rows for the reaction ordered are retrieved and listed in the file with the name specified by the user.

THE DATA TABLE FOR THE OBSERVATION SPACE SPECIFIED

This approach enables automatic composition of the table of data designed for further treatment. This table consists of data rows and columns. The latter define the conditions (further the term “factors” is used) to be specified to grant the reproducibility of the experimental measurements. Each factor involved can be represented by some formal (sequence number for the reaction, substituent, or solvent) or pithy (for instance the value of absolute temperature) specification. This is possible via specification of a set of distinguishable variable levels it may occupy (different substituent in a definite position, different solvents, etc.). Consequently, the table cited is comprised of rows, each of them representing some independently measured value called response (in the case under consideration the logarithm of rate or equilibrium constant) as well as of a set of corresponding factor levels for all variable factors involved. The factors stay constant for the entire observation space (*OS*) defined, can be specified for the table as a whole, and they can be omitted from particular data rows. So, the data table is represented by a set of the columns formed by response values. Each of them represents a selection of the levels for some given factor in correspondence with each data row specified. The table described represents actually the *OS* formed by the coordinate axes represented by factors and the response as a dependent value [7]. The latter can be considered a functional from the coordinates defined by the factor axes.

The retrieval system described enables automatic formation of differently defined particular experimentally described *OS* if the corresponding data are available in the database.

SOFTWARE FOR STATISTICAL TREATMENT OF THE DATA FROM SOME OBSERVATION SPACE FORMED

For some kind of conceptual quantitative interpretation of the particular *OS* the corresponding mathematical model has to be employed. For this purpose, the formal factor levels do not represent explicitly the arguments (descriptors) the response value is dependent on. Therefore, for each factor a single or several descriptor scales have to be defined and substituted for the corresponding selection of the factor levels. For some factors, the corresponding scale is independently available. For instance, the reverse value of the absolute temperature (1/*T*) can be substituted for each *T* value. This procedure is justified by the fundamental thermodynamic approach, although the empirical data obtained via the observation of the temperature dependence of the isobaric volume or the isochoric pressure for the ideal gas are usable for this purpose as well. The possible descriptor scales for the practically successful use are defined mostly on the basis of empirical data. Nevertheless, it is reasonable do distinguish between scales available independently of the particular *OS* to be interpreted, or established proceeding from the results of the (statistical) data processing applied to the corresponding data set itself. In the case of success, a new set of descriptors related to the influencing factors is generated.

In the field of reactivity, spectral and several other sets of data the approach called "Correlation Equations" (*CE*) has been widely used through years. As a result, different sets of substituent and solvent constants are available for the use as the descriptors related to corresponding factors (variable substituent or solvent). For these constants, the essential physical or chemical interpretations are assigned. One can differentiate the substituent polar or steric properties or the solvent polarity, polarizability, as well as the hydrogen bonding acidity and basicity, responsible for the existence of the specific solvation of the solutes.

The *CE* represent, in general, multiparameter linear functions of the response values on the sets of constants considered as descriptors. Each data column involved is characterized by the coefficients of the (multi)linear regression and characteristics that reflect the quality of description.

WAYS OF SOLVING THE PROBLEM OF NON-ORTHOGONALITY

The greatest difficulty connected with the multilinear approach is due to the non-orthogonality (mutual linear interdependence) of the descriptor scales used [8]. This is a case for both the descriptor scales defined independently of the data compilation processed and established as a result of this processing.

The classical way to define the scales of substituent or solvent constants is to find out a data set presumably related to a single definite property characteristic for some variable factor. If there exist a number of data series dependent on a single "pure" descriptor scale, the response values for all of them are in mutual (nearly) linear dependence. On the other hand, the observation of such kind of mutual linearity for a family of the sets of experimental data related to some common variable factor is considered a proof of the existence of such single scale. As the logarithms of the rate and equilibrium constants are linearly related to the free energy changes for reactions (equilibriums) or activation (rates), the term “linear free energy relationships” (*LFR*) has been introduced. As the spectral frequencies are linearly related to the corresponding excitation energies, the more general formulation “linear energy and free energy relationships” (*LEFR*) is used. It was very soon realized that these linear dependences are, in the general case, held roughly or not at all. This was considered as a proof that in the general case the *LEFR* do not hold (at least not precisely enough). As a further development of this approach, it has been assumed that more than a single property and the corresponding descriptor scale can be related to the same influencing factor. Therefore, multiparameter relationships were introduced. If more than a single factor is varied for some selection of data (e.g. the simultaneous variation of the substituent, solvent, and temperature), the multiparameter processing of data becomes inevitable.

In favourable cases, some descriptor scales may be nearly orthogonal for a definite selection of the corresponding factor levels. Nevertheless, even if this is the case for some more numerous set of data for different (combinations of) factor(s) levels, it is usually not true for several subsets of these data. If a number of points for the mutual plot of two independently defined scales related to some selection of data rows do not represent a corresponding linear dependence, this does not apply for the different subsets of data rows. For these this kind of interdependence may be more or less pronounced just by chance. In the limiting case, a straight line can be drawn through any two non-coinciding points.

For the case of the data matrixes with all positions filled, the method called Factor Analysis (*FA*) can be used to define a set of purely formal but orthogonal descriptors (called Factors) [9]. Up to now, there remain two main shortages of *FA*, which considerably limit its use. One of them is a lack of a fully distinct criterion for the detection of a number of significant factors. In addition, as already mentioned, its use is limited to the completely filled data matrixes, mostly not available for the sets of existing data. Although there may exist some ways for overcoming these difficulties, in this paper *FA* is not considered. The main reason is that this approach does not directly lead to the definition of the descriptors with some definite essential (physical or chemical) meaning.

Sufficiently numerous data compilations are usually incomplete. If for a large part of the points of the corresponding *OS* the response or the descriptor values are lacking (no empirical data are available), the non-orthogonality of the descriptor values for the sets of data rows for particular response columns is rather a common case.

Therefore, the quantitative characteristics of the degrees of orthogonality or non-orthogonality were defined and the procedure was elaborated for the derivation of the nearly orthogonal sets of the descriptor scales for the complete set of rows present in the data set processed. Besides the regression coefficients and their scaled versions, the contributions into the total variance (dispersion) by the different significant descriptors are characterized by their weights. These values differ from the corresponding orthogonal figures represented by the squares of the correlation coefficients (R^2_{xy}) between the column of the response values and of the corresponding descriptor column. The difference between this value and the weight obtained as a result of actual data processing is considered as a measure of the "mixed" part of the weight value. The corresponding scaled (by R^2_{xy}) value W_{mixd} represents the relative measure of the influence of the non-orthogonality on the contribution of the related descriptor. If the absolute value of W_{mixd} is larger than 1.0, the observed weight value is mainly defined by the concerted effect of other descriptors present and has little to do with the real influence of the descriptor it is nominally related to. Therefore, if $|W_{\mathrm{mixd}}| > 1.0$ (other limiting thresholds can be used), the corresponding term in the expression for the response value can be considered not to possess an essential meaning and deserves to be omitted.

A procedure was elaborated for removing (subtracting) from each subsequent descriptor scale the contributions of the preceding ones to obtain a set of nearly orthogonal "pure" (residual) descriptor vectors for the whole set of rows processed.

DATA PROCESSING PROGRAM

The software created for testing multilinear mathematical models for the description of a set of response columns related to some *OS* can be ordered for use [8, 10]. In this case, the main procedures are executed automatically as follows:

The regression of all response columns present in the table of initial data, with the selection for each of them of only such descriptors that are statistically significant on the risk level assigned, with the mixed component of weight not exceeding the threshold stated, with the weight exceeding the minimal limit value stated, and with the over-pumping effect (anomalously large standard deviation of the scaled regression coefficient at the expense of the mutual interrelation of descriptors) not exceeding the acceptable maximal value stated. Various other restrictions may be introduced if desired. The list of all results is calculated and output for every particular response column is obtained. The averaged results for the complete data set are output as well. The general contributions of the particular descriptors are characterized by their mean weights and degrees of representation (the parts (total numbers) of response columns that depend on the particular descriptors).

The whole set of response columns can be divided into two parts – the more and less precisely (roughly) described ones – via stating a corresponding criterion (e.g. the critical value of the multiple correlation coefficient). For these parts, all the averaged results are additionally output separately.

Several procedures are available for the ordering of the way the parallel (alternative) response values for a single data row for particular data columns have to be treated.

Different approaches are available for the elimination of strongly deviating response values. For the large set of different responses subjected to concerted processing, the traditional way of the use of the Student criterion on some risk level stated appears not to be a reasonable one. This is due to the different scaling and considerably varying precision of the description for the different response columns. Therefore, an alternative procedure is for use: for each data column from their roughly described set, the maximally deviating point is excluded until the transition of this column into the set of the more precisely described ones. Additional limiting conditions can be introduced via stating a maximal value of the excluded points for a single data column and a specified lower limit of the statistical degrees of freedom. This procedure opens a limited possibility for improving the quality of description at the expense of the eliminating of presumably unreliable data points.

There is a possibility of ordering the calculation of the missing values of descriptor rows, making use of the response values available for a corresponding row dependent on the related descriptor.

It is also possible to recalculate the values of the elements for some of the descriptors or for all of them retaining their scaling. If the suitable arbitrary initial approximations for these descriptors are substituted for the values of their elements, the set of them can be calculated in part or in whole from the very beginning. This is the most universal way for the (re-)establishing of the set of descriptors adjusted to some data matrix given. The nearly orthogonal set of residual descriptors obtained can be used for the calculation of the missing or unreliable values of the response columns.

Repeating the processing after the inclusion of these values, nearly orthogonal solutions for all response columns (at least for the ones from the more precisely described set) can be obtained. Simultaneously the predicted values are obtained for all respective positions in the data table for which the experimental values are missing or unreliable.

The justification of the inclusion of the particular descriptors into the set used for the description of some data table can be checked by the F-test of their significance for the whole set of data. This is possible if the averaged effect of the substitution of different selections of the (pseudo) random values for the descriptor columns tested is taken into account. The point is that, for a sufficiently numerous set of data series, the introduction of an additional “descriptor” consisting of random values appears to be formally statistically significant for some subset of response columns just by chance. For different selections of random values, the number of response columns for such subset remains nearly constant, but its concrete list may be a specific one for each new selection of random values.

For the data sets corresponding to multifactor *OS* the cross terms (products of two or more starting descriptor values for each row) can be automatically added. To avoid the additional over-pumping effect, caused by the large differences between the mean values and scale origins (zero points) for the descriptors multiplied in the course of the formation of cross terms, the possibility of the formation of these terms from the preliminary centred descriptor scales exists. For instance, a characteristic case is represented by the scale 1/*T*, because the ranges of temperature represented by the experimental data available lie mostly rather far from the absolute zero. The recalculation of the results to the ones corresponding to the initial non-centred scales is made possible, too. The technique of the formation of the cross terms can be used for the generation of the exponents of descriptors of different powers as well.

NONLINEAR MODELS

Besides the testing of multiparameter linear models, the testing and parameterization of nonlinear models is also possible. If the nonlinear function to be used is not included into the corresponding list of standard ones supplied, it can be additionally programmed and added for the additional compiling and linking. Another possibility is an additional EXE-file prepared by the user and conventionally named to grant calling of this function by the standard part of the computer program.

The numerical procedure of the generation of the set of partial derivatives to be used is executed automatically. If desired, the procedure of the generation of the analytical ones can be substituted for the numerical one. Techniques analogous to the one described for the nonlinear function can be used.

All problems, beginning with the non-orthogonality, that are encountered in the case of the multiparameter linear approach, arise in the case of nonlinear models for the set of the partial derivatives used in the iterative procedure. The corresponding difficulties can be overcome using the approach applied for the linear models. However, as a result of the higher complexity of the nonlinear problems, this cannot always be done completely successfully.

Several procedures are included for establishing reasonable initial approximations for the parameters of the processed nonlinear function. If desired, the user can add some additional version making use of the technique cited for the inclusion of the nonstandard version of the nonlinear function and analytical procedure for the calculation of partial derivatives.

SELECTION OF THE DIFFERENT VERSIONS OF THE CRITERIONS AND CONDITIONS

The fully automatic functioning of the whole system described is possible if all the (limiting) criterions, flags, and switches of the program are accepted on the default levels. In practice, this is rather unacceptable, and the corresponding initial data should be input additionally. Additional check of the results of the automatic formation of the table of responses and descriptors is rather reasonable, too. The choice between different data processing versions available and the criterions to be used is possible just by the assigning of the proper values to corresponding initial data (flags, switches, and the values of criterions preferred by the user). Acceptance of the default levels defined for them may be justified for the majority of them and their change would be reasonable only for advanced approaches to test some peculiar versions and ideas. Therefore, the list of the parameters, recommended not to be ”touched”, is defined.

The software is created in Fortran [11].

CONCLUSIONS

After statistical processing, the calculation of the values for all missing positions in the *OS* is possible. This means that the limit of the prognosis of new data, proceeding from it, could be achieved.

ACKNOWLEDGEMENTS

A part of this project was carried on using the funds of grants Nos 4590 and 6476 awarded by the Estonian Science Foundation, and the two year “long-term” grant from the Soros Foundation.

APPENDIX

LIST OF THE MAIN REACTION TYPES

1. Dissociation of hydrogen acids according to the Brönsted scheme
2. Rate constants of proton transfer
3. First-order nucleophilic substitution at nonaromatic centres, solvolysis
4. Second-order nucleophilic substitution at nonaromatic centres
5. Electrophilic substitution at nonaromatic centres
6. Addition to double and triple bonds
7. Elimination with the formation of a double or triple bond
8. Hydrolysis of carboxylic esters
9. Reactions of carbonyl compounds as electrophiles
10. Aromatic electrophilic substitution
11. Aromatic nucleophilic substitution
12. Tautomeric equilibria
13. Intramolecular rearrangements
14. Formation of complexes from general acids and general bases

REFERENCES

1. Palm, V. A. *Foundations of Quantitative Theory of Organic Reactions*. Khimia, Leningrad, 2nd ed., 1977 (in Russian).
2. Palm, V. A. (ed.) *Tables of Rate and Equilibrium Constants of Heterolytic Organic Reactions*, Vols 1–5. VINITI, Moscow, 1975–1979 (in Russian).
3. Palm, V. A. (ed.) *Tables of Rate and Equilibrium Constants of Heterolytic Organic Reactions*, Vols 1–6. University of Tartu, Tartu, 1984–1990 (in Russian).
4. Palm, V. Computer-managed automatic data retrieval and prognozing system for rate and equilibrium constants of organic reactions. *J. Chem. Inf. Comp. Sci.*, 1990, **30**, 409–412.
5. Palm, V., Jalas, A., Kiho, J. & Tenno, T. A computerized system for storage, processing and prognosis of data with orientation toward the use of correlation equations. *Org. React.*, 1997, **31**, 111–132.
6. Jalas, A., Kiho, J., Palm, V. & Tenno, T. Data structure and menu-based access of the rate and equilibrium constants of heterolytic organic reactions database. *Org. React.*, 1997, **31**, 135–140.
7. Palm, V. Some fundamental criteria of the scientific method and the internal structure of science. In *Estonian Studies in the History and Philosophy of Science* (Vihalemm, R., ed.). Kluwer Academic Publishers, Dordrecht, 2001, 91–110.
8. Palm, V. & Palm, N. On the total number and list of parameter scales significant for general quantitative description of solvent effects. *Org. React.*, 1993, **28**, 125–150.
9. Malinowski, E. R. *Factor Analysis in Chemistry*. J. Wiley & Sons, Toronto, New York, 1980.
10. Palm, V., Palm, N. & Tenno, T. Modification of data processing and interpretation of results related to the use of multiparameter correlation analysis. Introduction of additional characteristics and criterions. I. Application to the treatment of solvent effects. *J. Phys. Org. Chem.*, 2004, **17**, 876–889.
11. Microsoft Fortran PowerStation, Version 4.0, 0895 Part No. 65838.

Arvutistatud andmebaas ja tarkvara keemiliste reaktsioonide kiirus- ja tasakaalukonstantide otsimiseks, töötluseks ja prognoosimiseks

Viktor Palm, Natalia Palm ja Tiina Tenno

On loodud keemiliste reaktsioonide kiirus- ja tasakaalukonstantide arvutistatud andmebaas. Väljatöötatud tarkvara võimaldab leida andmeid ligi 500 000 reaktsioonide kiirus- ja tasakaalukonstanti sisaldavast andmebaasist. Andmete modifitseeritud statistilise töötlusega on võimalik prognoosida andmebaasis puuduvaid kiirus- ja tasakaalukonstante.