Proc. Estonian Acad. Sci. Chem., 2001, **50**, 3, 186–192
https://doi.org/10.3176/chem.2001.3.09

A COMPUTERIZED STORAGE AND RETRIEVAL SYSTEM OF pK_a VALUES OF HYDROGEN ACIDS

Alla JALAS, Viktor PALM, and Tiina TENNO

Institute of Chemical Physics, University of Tartu, Jakobi 2, 51014 Tartu, Estonia;
vpalm@ut.ee

Received 19 February 2001

Abstract. A computerized system was created and elaborated for information storage and retrieval of a database including information on approximately 20 000 organic and inorganic hydrogen acids. This paper describes the graphic menu system for a fast search of pK_a values of compounds on the basis of their chemical structure.

Key words: database, pK_a values.

INTRODUCTION

The database of rate and equilibrium constants of organic reactions has been published as the thesaurus *Tables of Rate and Equilibrium Constants of Heterolytic Organic Reactions* [1]. In parallel, a system for computerized storage of the database and a procedure for data retrieval were created [2–4]. By today the menu-based retrieval system for the computerized database has been completed and applied for the pK_a values of hydrogen acids[*]. This system is suited for a fast search of pK_a values of compounds on the basis of their chemical structure. The database contains approximately 200 000 pK_a values for over 20 000 organic and inorganic hydrogen acids in different media, compiled comprehensively and critically from original scientific journals from 1890 to 1990. For each compound, the data include pK_a values with error in logarithmic units, the solvent, and data on experimental conditions such as temperature,

[*] Complete description of all subtypes of hydrogen acids included in the pK_a database is available online [5].

pressure, the ionic strength of the solution, concentration of acidic and basic species etc., the experimental method, and references.

The following main types of hydrogen acids are presented in the pK_a database: OH-acids, NH-acids, amino acids, CH-acids, SH-acids, PH-acids, and tautomeric OH-, NH-, and SH-acids.

The pK_a database is supported by special programs and designed as a simple menu system similar to MS Windows.

Below a short description of the created data storage and retrieval systems is presented (see also [5]).

COMPUTERIZED STORAGE OF THE DATABASE

For the computerized storage of the database [1] a special coding system, the LINCS (Linear Coding of Structures), enabling the linear coding of the compounds and reactions, was created [6, 7]. The symbols applied for coding structural units (substituent or compound) are conventional symbols for denoting chemical elements (H, C, O, N, ... etc.), bond labels (**.** ordinary bond, = double bond, # triple bond, **:** aromatic bond), and separators (<, >). To simplify the coding, instead of the LINCS CODE the DIRECT CODES or macros of structural units (as a rule structural formulae) have been used. For instance, the LINCS CODE <<<C**.**H>**.**H>**.**H> of the methyl group CH_3 could be replaced by the DIRECT CODE CH3 or ME.

The reaction is encoded with indication of the breaking and forming bonds, notations, and locations of variable substituents and their bonds with other structural elements. For instance, the dissociation of carboxylic acid RCOOH $\leftrightarrow$ $RCOO^- + H^+$ (R stands for the label of a variable substituent) is encoded as <H[EY]O**.**CO**.**R>, where EY denotes the bond break EY $\rightarrow$ $E^+ + Y^-$ (E = H and Y = RCOO).

The coded data are input in the ASCII code, in one or several *initial file(s)* by using a conventional text editor. The encoded reaction type (reaction with a fixed and/or variable substituent(s)) and names of all variables occurring in the whole data set of the given reaction type (substituent(s), solvent(s), concentration, temperature, pressure) are shown in the head-lists TYPE and SUB. If there are such variables that remain constant for the whole set of data of the given reaction type, their names in the head-list SUB may be replaced by the constant value.

A further specification of the data will proceed in the head-list SER. It is possible to define the solvent and the values of those variables that remain constant for the given data subset following the head-list SER. Each information line following the head-list contains the specification of the variable(s), pK_a value, its error, an abbreviation of measuring method, the literature reference, and a note about the reaction conditions.

With the application of the PERFO program the *initial file* obtains numbered records and formally checked LINCS codes (a correct application of labels and

direct codes, regular placing of delimiters, etc.). These corrected initial files, the so-called *primary files*, which contain the corresponding LINCS codes for the reaction and the related information, form a *primary database*. The *primary files* are distributed into catalogues, whose names coincide with the extensions of the filenames. Through the primary database new data are added and existing data are corrected.

On the basis of the primary database a *searchable database* was formed by the conversion program DIANA. During the conversion the LINCS code of the reaction will be formed for each record of the *primary file*, supplementing the code given in the head-list TYPE with the data presented in the head-lists SUB and SER and those found in the information line. The LINCS code obtained does not contain any variables. The direct codes used in the code will now be replaced by corresponding LINCS codes from the TABLE OF DIRECT CODES. The obtained code contains only terminal labels. After that, the LINCS code will be replaced by an ograph (a graph with labelled nodes and arcs, where the nodes are the atoms and arcs are their chemical bonds). The related information is stored as the corresponding vectors of the parameters supplied with all the numerical data (the pK_a value with its error, temperature, concentration of measured compound, pressure, numerical code of solvent, composition of the mixed solvent, etc.), codes of reference, the experimental method used for measuring, and notes.

MENU-BASED RETRIEVAL SYSTEM WITH A GRAPHIC INPUT–OUTPUT SYSTEM

For data retrieval a graphic menu system named pK_aDATA was created [3, 4]. The aim of the system is to retrieve the data in the form of a menu, by moving, on the basis of the choices made, with retrieval orders from a general reaction type to a more and more concrete subtype to the final menu for compounds with definite structures.

The first menu of the pK_aDATA is a list of the most general types of hydrogen acids shown in the Introduction. When the sequence number of the desired general type of compounds is selected the next menu on the monitor screen will present a list of the subtypes of this general type, etc. In the final menu the substituents with their respective solvents attached will appear on the monitor screen. It is possible to select either a full set or a subset of lines (substituents with solvents) by selecting the respective number in the increasing order. Upon completion of the selection the number of lines chosen will appear on the monitor screen. If necessary, a selection of solvents may be carried out, in the case of which a numbered solvent list presented in the *s-file* will be displayed on the monitor screen. One menu choice brought to the monitor screen by means of the *m-* or *mm-files* (*mm-file* in the case of the final menu) consists of either one or several screen(s). It is possible to move between the prompted menu windows in both directions, forwards and backwards. After the final selection of the set of

compounds related to variable substituent(s) and solvent(s) user is completed, the search of the corresponding data starts.

The kernel of the retrieval system is a set of programs by means of which the reactions searched will be compared with the reactions in the database. Therefore, in order to prompt the retrieval order for the search procedure, the LINCS codes of the searchable reaction(s) have to be available. For the execution of the retrieval order the *l-file* corresponding to the final menu for compounds with definite structures is applied. The line(s) of the substituent(s) selected in the retrieval order will be taken from the *l-file* one after another, and the substituent varying in the general reaction code will be replaced by the LINCS code of the corresponding substituent.

In the described retrieval system each numeral characteristic of the name of the menufile (*m-* and *mm-files*) coincides with the serial number of the corresponding subtype of reaction given in the graphic menu. The names of *s-files* and *f-files* (the list of the database files, which are necessary for the search procedure of the compounds of the final menu) coincide with the name of the *mm-file*.

After completion of the search the information retrieved is presented on the monitor screen in the form of the OUTPUT TABLE. The data are presented by the solvents. At the top of the table, the types of the acid and solvent or solvent mixture are given. In the columns of the table are the substituent(s) or compound(s), the composition of the mixed solvent (as a rule, in mol per cent of one of the components), the temperature in °C, the pK_a value(s) with its (their) error(s) (in pK_a units), the experimental method (given in the list of the methods in the file METHODS), the reference(s) to the corresponding literature source (given in the list of the references in file REF), and notes. In notes ionic strength, pressure, and the concentration of the measured compound or different additive may be presented. More detailed notes are collected as footnotes to the table. Lists of the methods and source references follow the table(s).

All data of search are saved in temporary files denoted by p1, p2, etc. These files can be used with a special program CHEMOUT to display the data without repeating the search procedure. The tables on the monitor screen can be printed out.

The structure files for graphic representation of cyclic structures were prepared by the POLYFRGM program [2]. All subsidiary programs of the pK_aDATA system have been written in the programming language FORTRAN (MS FORTRAN Powerstation, 32-bit) in the DOS environment. All files necessary for the operation of the programs are input in the computer by using the ASCII codes.

The following intends to illustrate schematically a full session for the search of pK_a values of *p*-cresol and 4-*tert*-butylphenol in solvent H_2O–DMSO. The asterisks (*) denote the selections made by the user.

Screen-image No. 1.

Dissociation of Hydrogen Acids
1.* OH-acids
2. NH-acids
. . . . . .

Screen-image No. 2.

OH-acids
1.* Nonconjugated OH-acids
2. Conjugated OH-acids

Screen-image No. 3.

Nonconjugated OH-acids
1.* ROH
2. RCOOH
3. $R^1R^2P(O)OH$
. . . . . .

Screen-image No. 4.

ROH
1. Alcohols
2. Hydrated aldehydes RCH(OH)OH
. . . . . .
5.* Phenols

Screen-image No. 5.

Phenols
1. * Phenols with one substituent
2. Phenols with two substituents
. . . . . .

Screen-image No. 6.

Phenols with one substituent – RC_6H_4OH
1. R = H
2.* R = alkyl
3. R = functional group
. . . . . .

Screen-image No. 7.

Phenols substituted by the alkyl group – RC_6H_4OH
. . . . . .
3.* R = 4-Me (list of solvents)
. . . . . .
19.* R = 4-*tert*-butyl (list of solvents)

Screen-image No. 8.

Selected compounds:
Phenols substituted by the alkyl group – RC_6H_4OH
3.* R = 4-Me (list of solvents)
19.* R = 4-*tert*-butyl (list of solvents)
*All solvents ? (y/n) n**

Screen-image No. 9.

Solvents:
1. H_2O
. . . . . .
2.* H_2O–DMSO
(Several solvents can be selected)

Results of the search:

Dissociation of Hydrogen Acids

Phenols substituted by the alkyl group – RC_6H_4OH

Solvent: H_2O–DMSO

mol%* of DMSO	*t*°C	pK_a	± pK_a	Method	Ref	Notes
R = 4-Me						
2.72	20.0	10.40	–	NMR	H225	–
14.00	20.0	11.20	–	NMR	H225	–
. . . . .						
83.00	20.0	16.60	–	NMR	H225	–
R = 4-*tert*-butyl						
2.72	20.0	10.0	–	NMR	H225	–
14.00	20.0	11.20	–	NMR	H225	–
. . . . .						
83.00	20.0	16.50	–	NMR	H225	–

Methods

NMR – Nuclear Magnetic Resonance Spectroscopy

References

H225 – Halle, J.-C. & Schaal, R. *Bull. Soc. Chim. France*, **1972**, 3785.

ACKNOWLEDGEMENT

The authors acknowledge the financial support of this research by the Estonian Science Foundation (grant No. 3019).

REFERENCES

1. *Tables of Rate and Equilibrium Constants of Heterolytic Organic Reactions*. Palm, V. A. (ed.). VINITI, Vol. I–V, 1975–1979; Tartu University Publishing House, Suppl. Vol. I–VI, 1985–1990.
2. Palm, V. Computer managed automatic data retrieval and prognozing system for rate and equilibrium constants of organic reactions. *J. Chem. Inf. Comput. Sci.*, 1990, **30**, 409–412.
3. Palm, V., Jalas, A., Kiho, J. & Tenno, T. A computerised system for storage, processing and prognostication of data with orientation toward the use of correlation equations. *Org. React.*, 1997, **31**, 111–133.
4. 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.
5. Online. Internet, http:/www.chem.ut.ee/tktool/teadus/pkdb

6. Kiho, J. Linear coding of labelled graphs. *Org. React.*, 1970, **7**, 94–111 (in Russian).
7. Kiho, J. Principles of the treatment of graphs. *Proc. Comp. Centre Tartu State Univ.*, 1980, **45**, 83–89 (in Russian).

VESINIKHAPETE pK_a-ANDMEBAASI KOMPUTERISEERITUD SALVESTUS- JA OTSISÜSTEEM

Alla JALAS, Viktor PALM ja Tiina TENNO

Tartu Ülikooli keemiaosakonnas loodud pK_a-de andmebaas, mis sisaldab enam kui 20 000 orgaanilise ja anorgaanilise happe dissotsiatsioonikonstante, hõlmab kõiki tähtsamaid vesinikhappeid (OH-, NH-, CH-, SH-, PH- ja aminohappeid).

Artiklis on kirjeldatud andmebaasi komputeriseeritud salvestus- ja otsisüsteemi, mis võimaldab ühendi keemilise struktuuri järgi kiiresti saada pK_a väärtusi. Andmete sisestamis- ja otsisüsteemi aluseks on keemiliste struktuuride spetsiaalne kodeerimiskeel, nn. LINKS-keel. Salvestamisel transformeeritakse keemiline struktuur matemaatiliseks graafiks. Reaktsioonide otsimine andmebaasis toimub graafide võrdlemise teel. Andmete kättesaamiseks on välja töötatud lihtne graafiliste menüüde süsteem, mille abil saab leida ühendi(te) pK_a-de väärtuse(d), liikudes üldisemalt reaktsioonitüübilt konkreetsemale. Programm võimaldab menüüde vahel edasi-tagasi liikumist ja andmete otsimist viimases valikumenüüs toodud ühendite hulgast nii üksiku ühendi või mis tahes hulga ühendite kaupa. Kõik ühendid on esitatud lineaarsete või tsükliliste struktuurivalemite abil.

Andmed, s.o. konkreetne või üldistatud ühend ja sellele vastav asendusrühm, solvent, temperatuur, pK_a väärtus koos mõõtmisveaga ning kirjanduse viide tuuakse ekraanile tabelina, millele järgneb kirjanduse loetelu. Otsitut saab ekraanil edasi-tagasi vaadata. Tulemus on prinditav. Spetsiaalse programmi abil on võimalik varem tehtud valiku alusel saadud andmeid tuua uuesti ekraanile ilma lisaotsinguta.