SOME PROBLEMS OF CHEMICAL REACTIONS RETRIEVAL SYSTEMS

ZDZISŁAW S. HIPPE

Department of Computer Chemistry
University of Technology
6 Powstancow Warszawy Ave., 35-041 Rzeszow, Poland
zshippe@prz.rzeszow.pl

Abstract: The paper briefly discusses some organisational questions of distribution of chemical information over the computer network, emphasizing various tools available for chemical sciences, mainly: chemical reaction retrieval systems, systems for computer-assisted molecular modelling, then for computational chemistry, biocomputing, and information about human genes. From the above mentioned problems, the chemical reaction retrieval systems, like ORAC, REACCS, and SYNLIB are covered in details. Additionally, the latest philosophy of searching reaction databases, represented by family of tools from InfoChem GmbH (ChemReact41, CD-ROM — with over 2.5 mln of reactions — and STS), are critically reviewed.

1. Introduction

Dissemination of chemical information for chemical sciences and chemical industry through databased systems has been applied for many years. Already by the end of 60-ties, the Chemical Abstract Service (CAS) revolutionized the methodology of the development of subject index for abstracts of the primary literature, which directly lead to the first information database on chemical compounds. This database contained typical textual entries, like chemical formula, method(s) of preparation, basic physical and chemical properties, bibliographic data, etc. At the time of initializing the effort to develop such a base (CAC — database, and later the French database DARC) its role for the information society in foreseeable future was not fully perceived: the preliminary idea of the database architecture did not cover the capability of searching for hidden relationships, i.e. chemical reactions, logically connecting respective chemical compounds:

$$[\text{reagent(educt) (X)} \rightarrow \text{intermediate compound(s)} \rightarrow \text{product (Y)}]$$

Currently there are numerous institutions and organizations involved in longrange research devoted to acquisition, storage, processing (which may be called the creating of databases), and dissemination of information for chemical sciences. The dissemination of chemical information can be organized as the remote access to a specialized center (Chapter 2), or by the in-house access to local database (Chapter 3).

Detailed analysis of the available literature allows to draw a conclusion, that the creation of chemical information systems (database), from the very beginning has been directed towards scientific data on:

1) basic properties of chemical compounds,

2) specific properties of chemical compounds (e.g. molecular spectra) and

3) chemical reactions.

Simultaneously, some projects aimed at the elaboration of theoretical aspects of generating these types of data (mainly, selected properties of chemical com-pounds, molecular spectra, etc.) have been realized. Even partial success along this line may largely contribute towards creating the computer simulated data, without the necessity to gather it in a real experiment, frequently very expensive. However, the full realization of this idea seems to be rather a distant goal even for the most simple data type, i.e. physico-chemical properties. Development of synthetic spectra (all types of molecular spectroscopy, without restrictions as for the size and chemical character of compounds) creates distinct difficulties, whereas the prediction of organic reactions and organic syntheses — despite the engagement of many research teams worldwide — is still far from the satisfactory solution. This is probably the main reason of the high demand for real and reliable data for chemical sciences.

2. Databases for chemistry. The current status

Generally, an organized, fast, and reliable access to databases for uninterrupted exploitation of their content requires foundation of a specialized center equipped with necessary technical means, and human resources (staff) profoundly experienced in exploitation and maintenance of: a) selected databases, b) operating systems for dissemination of information and/or knowledge, and c) computer program systems used for generating data of various types. A model organization of this type is Computer-Assisted Organic Synthesis/Computer-Aided Molecular Modeling Center (CAOS/CAMM Center), established ~13 years ago at the Katholieke Universiteit Nijmegen (KUN), The Netherlands. CAOS/CAMM Center, a unique institution world-wide, has currently the most extended information resources for chemical sciences (databases and programs), hardware (worth roughly 4.5 mln US$), and networking facilities. Information resources are steadily available for about 200 group-users (research groups, research institutions, departments of chemistry in all Dutch universities, and also some selected industrial enterprises from the country). Additionally, CAOS/CAMM Center resources are accessible for two German research institutions, one Belgian, and two group-users from Poland (in Warsaw, and in Rzeszow). The demand for CAOS/CAMM services is very high, usually, there are roughly 43000 of individual on-line sessions a year. During the over 13-years long activity, two different groups of professionals have been created in the CAOS/CAMM Center, able to realize with profound expertise the following specific tasks:

task no. 1: optimum selecting of new databases and computer programs for chemistry; systematic training of end-users in remote accessing of various data-bases and programs; research advising; negotiations with new users; preparation of required documentation, e.g. manuals,

task no. 2: maintenance of database servers and computer program systems; organizing of reliable access services from all terminal nodes; updating of some databases (e.g. SwissProt or EMBL-New) by nights and creating working databases for in-house daily-use; inspecting and updating the status of databases and software licenses; requesting the service payment from users.

Databases and computer program systems available at CAOS/CAM Center have been gathered and grouped according to their function in teaching chemistry (at university level) or their role in chemical research. Thus, the following five groups of databases and programs may be enumerated:

Package A: chemical reaction retrieval systems, databases of molecular spectra, computer-assisted organic synthesis programs

Package B: computer-assisted molecular modeling

Package C: computational chemistry (quantum chemistry)

Package D: biocomputing (macromolecular sequence analysis of protein and nucleic acids)

Package E: information about human genes

It seems that information on resources of CAOS/CAMM Center may be extended with a comment about the statistics of the on-line accesses to various groups of databases and programs (Table 1), observed during the period of one year [Hippe, 1993]. From the data presented an interesting conclusion may be drawn that researchers using packages of utility programs for molecular modeling created the largest community of end-users. Similarly, numerous groups of people were employing research tools for the sequence analysis of amino acids in proteins and nucleic acids; this group of users was twice more active. It is believed that this finding points out that the heavy point of current research in chemistry is stepwise shifted towards these sub-fields of chemical sciences.

Table 1. Statistics of on-line accesses to information resources at CAOS/CAMM Center (1.01 – 31.12.1993)

Group (Package)	Number of group-users	Number of on-line sessions	% of sessions
A	23	~ 2 500	~ 5.9
B	65	10 666	24.9
C	22	5 333	12.4
D	55	23 996	55.9
E	10	400	0.9

Table 2. Statistics of on-line accesses to databases of chemical reactions and CASD programs at CAOS/CAMM Center (1.01 - 31.12.1993)

Access in group A to CASD programs (1-4) and databases (5-6)		% — globally	% — syntheses planing	% — databases of chem. reactions
(1) CAMEO	13	0.5	6.8	
(2) CHIRON	25	1.0	13.0	
(3) LHASA	156	6.2	80.2	
(4) SYNGEN	—	—	—	
(5) ORAC	2 022	80.9		87.7
(6) SYNLIB	284	11.4		12.3

Distribution of on-line accesses to databases of organic reactions and to programs for syntheses design (area of classic organic chemistry) is gathered in Table 2. The content of the table supplies an interesting observation: despite of the fact that databases of chemical reactions (speaking exactly, the *chemical reactions retrieval systems*) appeared on the market much later than computer-assisted synthesis design programs (CASD programs), already in 1993 the request for the access to databases containing information about *real* chemical reactions was many times larger, than that for programs (92.3% and 7.7%, respectively). One may speculate on reasons why computer-assisted reaction retrieval systems are more popular and widespread now? In the case of reaction retrieval systems the search of a database supplies immediate information on existing and well described, real chemical reactions. In the second case, however, results of processing — even by the most sophisticated CASD program — belong to the class of *computer simulated data* only. They have to be validated and verified against the literature; in many cases generated reactions may not exist in the real world of chemical transforms. Therefore, the present article deals mainly with the selected problems of retrieval of information about chemical reactions.

2.1. Databases of chemical reactions

The new type of databases of chemical reactions (correctly named *chemical reactions retrieval systems*) has been launched quite recently, i.e. in years 1986--1990 [Ott, 1996]. One distinct feature of the new databases generation is the inherent possibility to apply much more extended searching capabilities, well beyond the traditional mechanisms of text-oriented searching (i.e. searching for chemical names, names of reactions, bibliographic data, or keywords addressed to the content of the paper). Of paramount importance was the addition of the entirely new searching mechanisms: *a)* substructure searching, and *b)* searching based on atom-atom and bond-bond mapping. Both mechanisms require the access to special procedures enabling development of search queries in the form of structural questions, allowing retrieval information about structures and substructures. The option enabling to specify *any* substructure as a key for searching the database, supplies — as the end result — complete information about *all* reactions, in which the specified substructure may be contained, in starting chemicals (educts), in products, or simultaneously in both these species.

The mechanism of atom-atom mapping allows easy tracing of the rearrangements of key atoms during the reaction: additionally it may be applied as a powerful screening factor to distinctly decrease the information noise, accompanying the searching of any database.

Currently, the following chemical reactions retrieval systems are commonly used:

ORAC (***O****rganic* ***R****eactions* ***A****ccessed by* ***C****omputer*), contains:

• abstracts of articles from primary literature:	70 000 reactions
• abstracts of papers from secondary sources:	
– Comprehensive Heterocyclic Chemistry	38 000 reactions
– Theilheimer's Synthetic Methods	43 000 reactions

[resignation from the continuation of the database maintenance has been announced]

REACCS (***RE****action* ***ACC****ess* ***S****ystem*), contains:

• abstracts of articles from current literature:	
— general transforms	33 000 reactions
– reactions of organometallic compounds	10 000 reactions
– asymmetric reactions	11 000 reactions

- abstracts of papers from secondary sources:

– Comprehensive Heterocyclic Chemistry	35 000 reactions
– Theilheimer's Synthetic Methods	40 000 reactions
– Journal of Synthetic Methods	36 000 reactions
– Organic Synthesis	5 000 reactions

[development of a new software generation, the ISIS (Integrated Scientific Information System), has been announced. Also, a databased system IRDAS (ISIS Reaction Database Access System), modular client-server system with communication programs and powerful chemical graphics tools, for Windows environment (IBM, Macintosh), will be designed]

SYNLIB (***SYN***thesis ***LIB***rary), contains:

- abstracts of articles from primary literature: 96 000 reactions

[lack of atom-atom mapping capability, loss of results of the database searching (because the results are not automatically stored on the hard disk), large machine time of searching]

Recently, the very well known databases of properties of chemical compounds (resulting from the research activities at Chemical Abstract Service, Columbus, OH, and Beilstein Institute, Frankfurt/Main, are quickly converted to databases of chemical reactions. "Beta" versions of these bases, containing somewhat restricted number of reactions are already available. The first observations gathered during exploitation of these bases pointed out the necessity to develop entirely new searching algorithms, able to cope with huge number of facts (millions of reactions) and providing — as the result of searching — reasonably sized list of reactions, without information noise. It should be emphasized also that processing of currently known number of chemical reactions requires elaboration of a new class of algorithms, capable to generate *clusters of answers*, as it is, for example, done in the German system **CrossFire**.

One specific technical problem of accessing chemical reactions retrieval systems is connected with the required hardware resources. Namely, almost all reaction retrieval systems cannot be loaded and run on computers mainly used in this country, i.e. personal computers (PC's), under MS DOS or MS Windows. In fact, much more powerful computers are needed, controlled by Unix, Open VMS, or other operating systems. We may say that unfortunate direction of computerization accepted here some time ago filled our universities and research institutions with computer equipment, however, at the same time changed Poland into PC-land, where popular scientific computer programs, broadly available in Western countries, cannot be exploited unless the equipment would everywhere be replaced by workstations (SUN, SG or HP). Over two years long experience with a different approach, based on remote access to a distant machine in a specialized information center (CAOS/CAMM Center at KUN, Nijmegen, The Netherlands), has led to the conclusion about poor quality of our computer network (low speed of transmission, high package loss, poor round trip times, [Noordik, 1996]). In this situation — taking into account that even the fastest computer network may be "jammed" by graphics — the only solution of these contradictory problems seems to be the *local* implementation of **ChemReact41** on the personal computer, or on PC-machine in the library, either at university, university of technology, academy of medicine, or at any other research chemical institution. At this point of our discussion it should be emphasized, that **ChemReact41** (jointly with **CD-React**) is the single chemical reaction database that may be exploited using personal computers of the type IBM-PC or Macintosh [Loew, et all., 1997]. This database is described in details in the next Chapter.

3. Databases for chemistry. Perspectives of application

The initiative to develop the database environment of chemical reactions began many years ago, and resulted from the direct cooperation of ZIC (Berlin) with VINITI (Moscow). Both organizations have jointly registered in the time span 1975–1991 roughly 3.3 mln chemical structures and facts, creating a so called *structural database*. Soon after reunification of Germany and establishing ZIC GmbH in Berlin, from this structural database a reaction database of *2.5 mln organic reactions* has been derived by the ZIC-Institute. As a commercial form of the 2.5 mln reactions file, the **InfoChem Reaction Database** has been generated, being now a source for other databases, among them — for example **ChemReact41**. It is believed, that chemical reactions described in the primary literature after 1991 will be registered and loaded in foreseeable future.

ChemReact41 represents a true chemical reaction retrieval system of the new generation. Reaction databases of the older type were equipped with text-oriented searching mechanisms only (searching names of compounds, names of reactions, bibliographic data and/or keywords addressed to chemical content of a publication). In ChemReact41 we have additionally two other mechanisms of searching: *a) searching for substructures*, and *b)* based on an atom-atom mapping, a *reaction substructure searching*.

The ChemReact41 database has been developed using the concept of *reaction type*. Reaction type may be defined by means of a set of reaction centers (reaction sites) and their meaningful molecular surroundings. Two (or more) chemical reactions, having the same reaction centers and the same molecular surroundings, belong to the same type. For example, esterification reactions of benzoic acid with methanol, ethanol, propanol, isopropanol, etc., are components of the set of reactions of the same type. Therefore, ChemReact41 is a database of reaction types, thus not an ordinary database of organic reactions. The resources of ChemReact41 are as follows:

41 300	chemical reactions (11% from patent literature),
76 000	structures,

26 000 literature references (56 professional journals abstracted regularly, 30% references from patent literature).

ChemReact41 contains the "best" example of a given reaction type, if the reaction is referred in the literature at least 5 times, when one of these citations is contained in an internationally accepted organic synthesis journal, and at least one of the described examples has yield higher than 50%. Each reaction kept in the database is associated with complete bibliographic data, short comment about conditions of experiment, yield, stereochemical information about the reaction, and a brief information about molecular spectra (if any) in the relevant article. One unique feature of ChemReact41 environment is the application of dynamic data exchange mechanism (DDE — mechanism) for navigation between the basic module and the parent Reaction Database, frequently called **InfoChem CD-ROM**, containing over *2.5 mln* of chemical reactions. Thus, the results of a query in the base of reaction types (ChemReact41) may be — after completion — transferred with the access software CD-React to the parent base in order to display the additional examples available for the reaction type regarded. In this situation, the capacity of the hard disk should exceed 1.5 GB.

The described program environment of chemical reactions consists of the following programs:

PC-Search and *CD-React Access Software*,

having the access to databases:

ChemReact41 (41 330 reactions) and *InfoChem CD-ROM* (2 500 000 reactions)

respectively. Additionally, "front-end" type programming tool is necessary to create "structural" queries. This tool is not supplied by InfoChem GmbH; it may be downloaded, as shareware, through Internet from two sources:

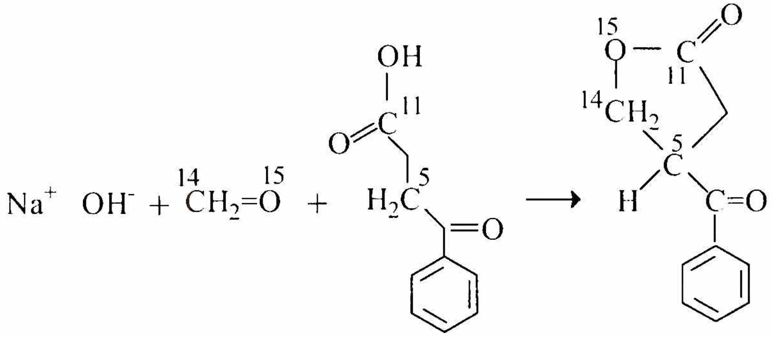
MDL Information Systems, Inc. (computer program *ISIS/Draw*)

or

STN International (computer program *STN Express*).

Structural queries (about a given structure, substructure, reaction) are generated by means of ISIS/Draw or STN Express and transferred — via Windows clipboard – – to PC-Search. The result of a query in the ChemReact41 database is a hit list, containing either *list of reactions* or *list of molecules*, according to initial requirements of the user. Pairs of lists may be combined using basic logical operators (AND, OR, NOT). Besides, any list of molecules may be combined with the information about the role of a particular structure (substructure) in retrieved reactions (for example, as reagent, product, catalyst, solvent, etc.). Queries of the type "data" (yield, author(s), journal, year of publication, language of the article, reaction conditions, reaction type, identification number of the reaction, etc.) may be directly inputted to PC-Search.

One characteristic feature of the discussed database environment is almost successful avoidance of redundancy and unnecessary repetitions of reactions – – a phenomenon being the Achilles' heel of other databases. For example, the reaction described by Cignarella [Cinarella, 1980]:



is registered in ChemReact41 database as reaction type 328673, and can be found using properly fixed search keys. The DDE-mechanism transfers the search process to the parent database (InfoChem CD-ROM), discovering reactions of the same type, published three times by the same team of authors (see [Shams, et all. 1986]). The first publication is in Polish Journal of Chemistry; it is also repeated in Egyptian Journal of Chemistry, and then again in Review of Roum. Chim. Thus the main database (ChemReact41, the database of reaction types) contains only *one example* for the reaction type searched. The selection of the paper published by Cignarella and coworkers is due to a higher significance of the Italian journal (impact factor), in comparison with the Polish, Egyptian, or Romanian journals, respectively.

The InfoChem Reaction Database product line (for Windows) has been recently enriched by Synthesis Tree Search (**STS**) program to retrieve all published *syntheses* for a given target molecule or all *reactions* using a given molecule as a starting material. In this way **STS** can be used for bi-directional navigation throughout the InfoChem database on CD-ROM, enabling easy development of *Synthesis Tree* (i.e. all synthesis steps leading to the target) or *Reaction Tree* (i.e. all known reactions starting from the selected molecule). In the first case, each reactant may be used as a new target, whereas in the second case, each product may be in turn used as a new starting material, allowing to display all reactions for that chemical.

4. Conclusions

The present article informs very briefly about the situation (in this country) in the field of databases of chemical reactions. Also, a specific type of databases has been examined. It may be expected, that computer-assisted chemical reactions retrieval systems will — for many reasons — have soon dominating influence in chemical research. This view may be substantiated by many reasons. Three of them seemed to be worth mentioning:

- the general tendency to abandon passive knowledge (contained in books), with a shift to new forms of information technology, enhanced by direct use of similarities, case-based inferencing, or analogies;
- the general tendency — almost worldwide — to stepwise decrease funds available for scientific research. In this situation, the access to fast, profound, complex, dedicated, and cheap information that can be mined by *reaction retrieval systems* is a very promising alternative, and
- the general tendency — in current and future research in chemistry — to shift their heavy point to problems of finding (synthesizing) compounds with a priori specified properties, e.g. drugs, herbicides, substances increasing the growth of plants (worldwide problem of food supply for people), catalysts (for various processes), inhibitors (for various reactions), new materials, etc. Making these types of investigations, overlooking of any related reaction must be strictly avoided. For this reason, the access to a good, well designed database of chemical reactions is of highest priority.

References:

[Cignarella, et all., 1980]

Cignarella G., Grella G., and Curzu M. M.: *Hydroxymethylierung von 3-Aroyl propionsuren; Verbesserte Synthese von 4-Aroyl-2(3H)-dihydrofuranonen und ein leichter Zugang zu 3-Aroyl-but-3-ensuren.*
Synthesis **10** (1980) 225-228

[Hippe, 1993]

Hippe Z. S.: *Perspectives and Applications of Computers in Chemistry as Affected by Growth of Information Technology.*
Report 3510PL92240, Katholieke Universiteit, Nijmegen 1993

[Loew, et all., 1997]

Loew P., Saller H. ,Hippe Z. S., and Nowak G.: *InfoChem GmbH Program Environment in the Search for Knowledge on Chemical Reactions.*
in: Hippe Z. S., and Ugi I. K. (Eds.): *MultiComponent Reactions and Combinatorial Chemistry*,
University of Technology Editorial Office, Rzeszow 1997, pp. 96-107

[Noordik, 1996]

Noordik J. H.: *Private information.*

[Ott, 1996]

Ott M. A.: *Computer Methods in Synthesis Analysis. Application to Reaction etrieval*

and Syntheses Design.
PhD-Thesis, Katholieke Universiteit, Nijmegen 1996.

[Shams, et all., 1986]

Shams N.A., Hamed A.A., Donia S.G., Shaker S.A.: *Synthesis of 4-aroyl-4,5-dihydro 2(3H)-furanones and their reactions with some nucleophiles.*
Polish J. Chem. **60** (1986) 143-149

Shams N.A., Hamed A.A., Donia S.G., Shaker S.A.: *Synthesis of 4-aroyl-4,5-dihydro 2(3H)-furanones and their reactions with some nucleophiles.*
Egypt. J. Chem. **29** (1986) 165-172

Shams N.A., Hamed A.A., Donia S.G., Shaker S.A.: *Synthesis of 4-aroyl-4,5- dihydro-2(3,H)-furanones and their reactions with some nucleophiles,*
Rev. Roum. Chim. **31** (1986) 541-546.