# Status and Future Developments of Reaction Databases and Online Retrieval Systems

ANDREAS BARTH

STN International, FIZ Karlsruhe, D-7500 Karlsruhe, FRG

Received May 28, 1990

In this paper the status of reaction databases and online reaction retrieval capabilities are outlined. The major sources of online reaction information are the CA and Registry Files from Chemical Abstracts, together with the Beilstein Database of organic chemistry. Since none of these files is a true reaction database, the search capabilities are rather limited. CASREACT provides an extension to these files with more systematic indexing of reactions. Today, the online hosts do not provide much support of reaction retrieval. Several features which are essential for a reaction data service are discussed in this paper. A support of stereochemistry, a reaction site searching, a multistep reaction retrieval, and the search for similar reactions are the major features to be developed.

## INTRODUCTION

The requirements for a reaction documentation and reaction retrieval system have been reported in the literature for a long time.[1-18] Several in-house systems have been developed to satisfy the urgent documentation needs of the chemists. Two of these systems, REACCS[19-25] and ORAC,[23-26] are commercially available and have been installed worldwide more than a hundred times. In the online business, no system with comparable capabilities is available today. However, there are several sources of reaction information, i.e., the CAS ONLINE databases[27-31] or the Beilstein file,[32,33] but the documentation is rather limited and the retrieval capabilities that are offered by the online hosts offer little support of reactions. Since 1987 Chemical Abstracts Service has extended its set of files and offers now a reaction database called CASREACT. This database covers the chemical literature since 1985 and contains the complete reaction scheme including both single-step and multistep reactions. According to the concept of Chemical Abstracts Service, this database is based on a previous search in the CAS Registry File, e.g., a substructure search. Although, this provides more support of reaction retrieval, it does not cover the more general requirements of most synthetic chemists in the laboratory.

It is the focus of this paper to analyze the current status of online reaction databases and the capabilities of the current retrieval systems and outline the future requirements for online systems. In the first part of this paper, the present status of the online databases and retrieval systems is summarized. In the second part, a discussion of the requirements which are not available today through the online hosts for online reaction retrieval is presented.

## REACTION DATABASES AND REACTION RETRIEVAL SYSTEMS

**Available Sources of Reaction Information.** There are several sources of reaction information available both as in-house files or through online hosts. Not all databases are reaction files, some of them have a broader focus like the CAS ONLINE databases, CA and Registry or Beilstein Database. In Table I the available reaction databases are listed together with the number of reactions and the coverage. All figures in Table I except those for CASREACT and CRDS have been taken from ref 34.

The databases of Table I can be classified according to their coverage as complete, selective, specific, or topical. While CASREACT is a general database covering the complete organic chemical literature since 1985, the Current Literature File is focusing on topicality. The *Journal of Synthetic Methods* and the *Organic Synthesis* files are more selective,

**Table I.** Available Chemical Reaction Databases[34]

| database | no. of reactions | coverage | available through |
|---|---|---|---|
| CASREACT | 750 000[a] | ≥1985 | STN |
| CHIRAS Asymmetric Synthesis | 5 000 | ≥1975 | REACCS |
| Current Literature File (CLF) | 25 000 | ≥1983 | REACCS |
| Journal of Synthetic Methods | 29 000 | ≥1980 | REACCS |
| ORAC Academic Collaboration | 7 500 | ≥1987 | ORAC |
| ORAC Core Database | 50 000 | ≥1900 | ORAC |
| ORAC Heterocyclic | 15 000 | ≥1980 | ORAC |
| Organic Synthesis | 5 000 | ≥1921 | REACCS |
| Chemical Reactions Documentation Service (CRDS) | 79 000 | ≥1942 | Maxwell[35,36] |
| Theilheimer/Synthetic Methods of Organic Chemistry | 42 000 | 1946-1980 | ORAC, REACCS |

[a] The file contains about 65 000 documents with more than 750 000 single-step and 900 000 multistep reactions. There are approximately 200 000 reactions added per year.

**Table II.** Chemical Reaction Databases in Preparation

| database | no. of reactions | increase per year | coverage | available through |
|---|---|---|---|---|
| Beilstein Reactions[38] | >5 000 000 | 200 000 | ≥1830 | STN |
| ChemInform[40-41] | 170 000 | 70 000 | ≥1988 | STN, ORAC, REACCS |
| ZIC Reactions[39] | 1 600 000 | 250 000 | ≥1980 | STN, ORAC, REACCS |

and CHIRAS and the Heterocyclic File cover a specific area of organic chemistry. The database CRDS[35,36] is the oldest database with reaction information, and it covers novel organic chemical reactions, reagents, synthetic methods, and other data of interest to the experimentalist. It includes Theilheimer's *Synthetic Methods of Organic Chemistry* and the *Journal of Synthetic Methods*. The in-house version available for REACCS and ORAC is only a subset of the online version. Although there are no adequate search features available for the online version, it is an important database since it is still the only publicly available online source for intellectually selected reactions. For a more detailed description of these databases see refs 34, 36, and 37.

In Table II three new chemical reaction databases that are currently in preparation are listed. Beilstein Reactions[38] is a new database under development based essentially on the reaction information in the current Beilstein File. Both the Beilstein[38] and the ZIC[39] reaction files are rather huge with respect to the number of reactions. The ChemInform[40,41] database is a medium-size file with a selective coverage of the literature being developed for both in-house and online users. In the following, the status quo of reaction information is discussed based on the CAS ONLINE system, including CA, REGISTRY, and CASREACT files, and the Beilstein Da-

0095-2338/90/1630-0384$02.50/0 

```
REGISTRY File:

=> s abietic(w)acid
           216 ABIETIC
       2637752 ACID
L1         209 ABIETIC(W)ACID


CA File:

=> s l1/p
L2         185 L1/P

=> d 3 ti au so py hit

L2    ANSWER 3 OF 185
COPYRIGHT (C) 1990 AMERICAN CHEMICAL SOCIETY

TI    Photooxidation of resin acids
AU    Gigante, Barbara; Marcelo-Curto, M. Joao; Lobo, Ana M.;
      Prabhakar, Sundaresan; Slawin, Alexandra J.; Rzepa, Henry S.;
      Williams, David J.
SO    J. Nat. Prod., 52(1), 85-94
PY    1989
IT    ***60188-95-6P***
         (prepn. and acetylation of)
IT    ***17751-36-9P***
         (prepn. and photooxidn. of)
IT    5335-63-7P   ***53655-46-2P***   53655-47-3P   57706-50-0P
      123887-69-4P   123887-71-8P   123887-72-9P   123887-73-0P
      (prepn. of)
```

**Figure 1.** Example of a search for substance preparation.

tabase in its current implementation.

**CAS ONLINE System.** The CAS ONLINE system[25,27-31] consists of several databases with different data structures covering various subjects. The basis of this system are the CA and CAOLD Files with the bibliographic information of the chemical literature and the REGISTRY File containing more than 10 million chemical substances. The latter is a substance database with the CAS Registry Number (RN) as the primary key. All chemical substances which have been quoted in the literature after 1957 are indexed in the REGISTRY File. A search of the topological structure or substructure is the main access point for this database. In addition, there are several fields which allow a search of so-called dictionary terms, e.g., chemical names, element counts, and molecular formulas. The CA File comprises the complete bibliographic information of the chemical literature since 1967, and CAOLD contains the literature partly since 1957. They are both bibliographic files with the CA Abstract Number (AN) as the primary key. A cross-over between CA and REGISTRY can be performed with the registry number (RN). Hence, it is possible to search for a class of substances in the REGISTRY File, take the results into the CA File, and display the references of interest.

There are limited possibilities to search for chemical reactions in the CA/REGISTRY system.[42] From the structure of the two databases, it is clear that the reaction searches are based on the individual substances participating in a chemical reaction. The logic for such a search is as follows:

1. Build the structure query.
2. Perform a substructure search.
3. Display the answer(s) in the REGISTRY File.
4. Extract the relevant answer(s), i.e., the registry numbers.
5. Display the latest literature references from CA.

The first two steps could be replaced by a dictionary search, especially in the case of inorganic compounds, but the overall strategy is the same. There are basically two aspects which can be searched in the CA/REGISTRY system: (1) preparation of substances and (2) reactions from A to B. The first step of a reaction search in the CA/REGISTRY system is a search for the chemical substances involved in the reaction. To find substance preparations, one could start with a (sub)structure search in the REGISTRY File. In the example in Figure 1, a search for abietic acid is performed in the REGISTRY File, and the answer is then crossed over to the CA File. Here, it is qualified as a preparation in the search. The final result is restricted to those substances for which a preparation method is described in the literature. As a result, one obtains the literature references. An extract of such a document is also shown in Figure 1.

In the next example, a search for the reaction (transformation) from A to B is requested. At first, one has to search for the substructures I and II (Figure 2). Let us assume that we know the CAS Registry Numbers in the parent compounds. In this case we can build the structures using the CAS Registry Numbers in the STRUCTURE command. After the creation of the structure, all hydrogen atoms are removed, and they are considered as free sites in the subsequent substructure search. One obtains 103 substances with pyrrolizidinedione and 62 with the pyrrolizinone skeleton. These answer sets are crossed over to the CA File. Here, it is requested that both structures (L2 and L4) are found in the same document together with the keyword "reaction". (Instead of "reaction" one could also use a more specific keyword like "photodecomposition" or "photolysis".) As an example, the reduction of pyrrolizidinediones to pyrrolizinones is retrieved. Two documents are found, the title (TI), graphic information (GI), and abstract (AB) are displayed.

**Beilstein Database(s).** The Beilstein Database[32,33] has been available online through STN International since December 1988 and on DIALOG since November 1989. It comprises the structures and factual data of approximately 3.4 million organic substances covering the literature period from 1830 to 1959 from Beilstein's *Handbook of Organic Chemistry*, and from 1960 to 1979 with excerpts from the original literature. While the data from the Handbook has been critically reviewed by the editors of the Beilstein Institute, no checking has been done for the literature excerpts. The scope of information may cover substance identification, including the chemical structure, general information, physical properties, and chemical data—including reaction information. For the Handbook

**REGISTRY File:**

```
=> str 18356-28-0
:end
L3    STRUCTURE CREATED

=> s l3 sss full
              :
L4            103 SEA SSS FUL L3

=> str 98216-93-4
:end
L5    STRUCTURE CREATED

=> s l5 sss ful
              :
L6             67 SEA SSS FUL L5
```

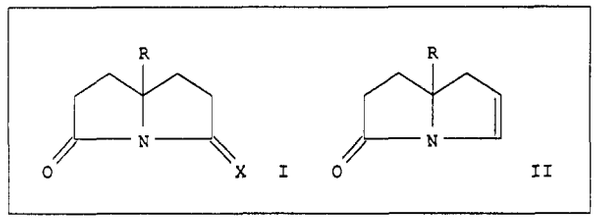
**CA File:**

```
=> s l4 and l6 and reaction
             34 L4
             10 L6
         765480 REACTION
L7            2 L4 AND L6 AND REACTION

=> d 1 an ti abs hit

L7    ANSWER 1 OF 2
COPYRIGHT (C) 1990 AMERICAN CHEMICAL SOCIETY

AN    CA111(17):153556w
TI    Reaction of novel imide reducing reagents with
      pyrrolizidinediones
GI
```




```
AB    The redn. of pyrrolizidinediones I (R = H, Me; X = O) with
      Dibal and LiBHEt3 affords the alcs. I (X = H,HO) in good
      yield. LiBHEt3 also reduces N-methylglutarimide in 53% yiel
      The combination of NaBH4/MeOH/Ac2O/CH2Cl2 selectively reduc
      an imide in the presence of an ester. Hexahydro-5-
      (methylthio)-3H-pyrrolizin-3-ones are products of the NaBH4
      redn. of pyrrolizidinediones in MeSH/CH2Cl2/Ac2O. I (R= H,
      Me, CH2CH2CO2Me; X = H,MeS) and I (X = H,HO), are
      intermediates in the synthesis of pyrrolizinones II. The
      structure of I (R = Me, X = H,HO) was detd. by x-ray
      crystallog.
```

**Figure 2.** Example of a chemical reaction search.

period the factual data are stored in the database together with the literature references, while the unchecked data consists, in most cases, only of the literature references. There are three classes of chemical reaction data which can be found in the Beilstein database:

- preparations (86% of substances)
- chemical behavior (13% of substances)
- biosynthesis and natural products (1% of substances)

The search of these different types of reaction information will be illustrated with some examples. A search for the preparation of a substance can be performed in a similar way as in the CAS ONLINE system. Title compounds, i.e., registered substances, can be searched as (sub)structures. In Figure 3 an example is given for the search of preparation methods of adrenalin compounds. Here, the search for the name segment "adrenalin" in the field chemical name segment (CNS) is combined with a search for the availability (presence) of preparation information in the document. The result comprises seven derivatives of adrenalin. The display shows the substance data of the title compound (adrenalin) and one reference for the preparation.

In our next example, the focus of interest is on the chemical behavior of sydnone derivatives. In Figure 4, a search for the name segment "sydnon" in the chemical name segment (CNS) field is combined with the availability of reaction information. The result comprises the chemical behavior of 50 different sydnone compounds one of which is displayed in Figure 4.

Information on biosynthesis can be searched in the field isolation of natural products (INP). In the example, a search for the isolation of brucine derivatives from strychnine compounds is performed (Figure 5). A name search has been performed instead of a substructure search to simplify the

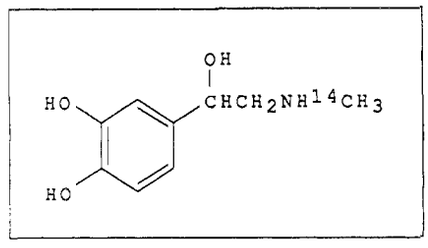
```
BEILSTEIN File:

=> s adrenalin/cns and pre/fa
            10 ADRENALIN/CNS
       2619028 PRE/FA
L8            7 ADRENALIN/CNS AND PRE/FA

=> d 4

L8    ANSWER 4 OF 7

BRN   2734195  Beilstein
MF    C9 H13 N O3
SY    ***Adrenalin***
FW    183.21
SO    5-13
LN    15336; 2817
RN    93117-67-0; 115861-64-8
```




```
Preparation:
PRE
      Reference(s):
      1. Pichat, Audinot, Bull.Soc.Chim.Fr., <1961>, 2256, CODEN:
      BSCFAS
```

**Figure 3.** Search for information about the preparation of adrenalin compounds.

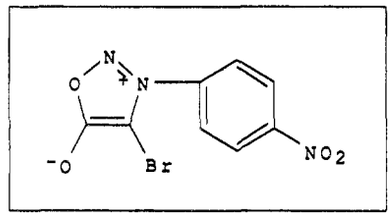
```
BEILSTEIN File:

=> s sydnon/cns and rea/fa
          328 SYDNON/CNS
       377608 REA/FA
L9          50 SYDNON/CNS AND REA/FA

=> d 4

L9    ANSWER 4 OF 50

BRN   1626712  Beilstein
MF    C8 H4 Br N3 O4
SY    ***4-Brom-3-(p-nitrophenyl)-sydnon***
FW    286.04
SO    5-27
LN    32053; 16540
RN    14606-04-3
```




```
Chemical Reaction:
REA
      Reference(s):
      1. Puranik, Suschitzky, J.Chem.Soc.C, <1967>, 10, 1006,
      CODEN: JSOOAX
```

**Figure 4.** Search for the reactions of sydnone compounds.

procedure. In the first answer, one finds a reference describing the kinetics of the reaction.

It is also possible to perform a search on the chemical names of the compounds in a reaction. These names are indexed in subfields of preparation (PRE.xxx) or reaction (REA.xxx). In Figure 6 the reaction scheme for preparation is shown. Here, the substance C is built from compounds A and B, and substance D may be a byproduct. The names of the starting

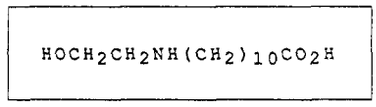
```
BEILSTEIN File:

=> s brucin?/cn and strych?/inp
            10 BRUCIN?/CN
           211 STRYCH?/INP
L10           4 BRUCIN?/CN AND STRYCH?/INP

=> d

L10   ANSWER 1 OF 4

BRN   1094415  Beilstein
MF    C23 H26 N2 O4
SY    ***Brucin***
FW    394.47
SO    5-27
LN    32183; 289
```




```
Isolation from Natural Product:
INP   Kinetik der Extraktion aus Strychni Samen
      Reference(s):
      1. Okada,Kawashima, Yakugaku Zasshi, 89, <1969>, 1345, CODEN:
         YKKZAJ
         CA: 15705, 72, 1970
```

**Figure 5.** Search for the isolation of brucine derivatives from strychnine compounds.

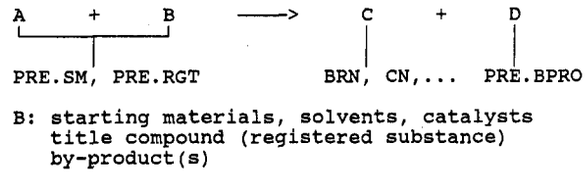



**Figure 6.** Reaction scheme for substance preparation.

materials are indexed in the field preparation starting materials (PRE.SM) or preparation reagents (PRE.RGT), and the title compound can be identified by the Beilstein Registry Number (BRN), the chemical name (CN), or the structure (STR).

It can be stated that the Beilstein Database is a substance-oriented factual database containing a lot of different reaction information. It is clear, however, that these substance searches that are based on nomenclature are not sufficient for the requests of the synthetic chemists. The standards and conventions for nomenclature have changed over the years. For most substances from the period 1960–1979 names are not available in the database. Furthermore, nomenclature searches are not a substitute for a topological search. The development of a true Beilstein reaction database will be an important extension of the current reaction information sources. This new file will be based on reactions instead of substances, including the full structures of the starting materials and the products as well as the corresponding reaction sites.

**CASREACT File.** The CASREACT[37,43,44] provides an extension to the limited possibilities of the CA and REGISTRY Files with respect to reaction searching. CASREACT is a document-based file containing reaction information described in more than 100 journals of the chemical literature since 1985. It covers approximately 65 000 documents containing more than 750 000 single-step and about 900 000 multistep reactions. The primary key is the CA Abstract Number (AN), and the CA Registry Number is indexed in various fields associated with the role of the substances in a chemical reaction. Each substance can be classified according to its role in the reaction process as reactant, product, reagent, solvent, and catalyst. Again, the first step is a substructure search in the REGISTRY File to obtain the list(s) of registry numbers. In the example in Figure 7, the results from the substance search have been crossed over to CASREACT, and the substances of the answer sets L5 and L6 are classified as reactants (RCT) and products (PRO), respectively. In addition, the search has been restricted to a yield which is greater than 50. The display shows the reaction scheme of the first answer. If the bibliographic information is also required, then this information can be obtained via a direct access to the CA File without a file change.

**Comparison of Reaction Searching in CAS ONLINE and Beilstein Database.** Both the CAS system and the Beilstein Database provide various possibilities to search reaction information. A comparison of these databases with respect to reaction information is summarized in Table III. It is possible to search for reaction information in the CA and Beilstein Databases. However, both files are lacking a systematic indexing of reaction information. In the CAS Registry System the stereochemistry is only recorded as textual terms. However, the REGISTRY File is currently upgraded to include the stereochemical information in topological form for both display and search. The Beilstein Database contains stereochemical information for display, but this is currently only available on DIALOG. The CASREACT File is an extension of the CA File, providing a more systematic indexing with respect to reactions as well as better search capabilities, e.g., the number of reaction steps, the yield, or the role codes. The restricted reaction retrieval, however, which is based on substructure searches on the individual compounds participating in the reaction is a major disadvantage of this database. Although there are some queries which can be answered with these databases, it is clear that none of the existing online databases provides sufficient reaction information. In addition, the retrieval systems offer good capabilities for text and substructure searching, but no system currently offers a true reaction searching.

## REQUIREMENTS FOR THE RETRIEVAL OF CHEMICAL REACTIONS

**General.** It is useful to outline the features required for an online retrieval of chemical reactions. The most important requirements for a reaction retrieval system are[14,37,45]

```
REGISTRY File:

=> s phenyl(w)hydrazon? and casreact/lc
       3277664 PHENYL
        107400 HYDRAZON?
         57877 PHENYL(W)HYDRAZON?
        897288 CASREACT/LC
L11       2720 PHENYL(W)HYDRAZON? AND CASREACT/LC

=> s methyl(w)indol? and casreact/lc
       5086933 METHYL
        193921 INDOL?
          1363 METHYL(W)INDOL?
        897288 CASREACT/LC
L12        229 METHYL(W)INDOL? AND CASREACT/LC


CASREACT File:

=> s l11/rct (l) l12/pro and yd > 50
           545 L11/RCT
           182 L12/PRO
             5 L11/RCT (L) L12/PRO
         35526 YD > 50
L13          3 L11/RCT (L) L12/PRO AND YD > 50

=> d 2 an ti au so py fsam

L13   ANSWER 2 OF 3
COPYRIGHT (C) 1990 AMERICAN CHEMICAL SOCIETY

AN    110:74549  CASREACT
TI    Role of phosphorus adducts in the indolization reaction
      between arylhydrazones and phosphorus trichloride
AU    Baccolini, Graziano; Dalpozzo, Renato; Errani, Ermanno
SO    Tetrahedron, 43(12), 2755-60
PY    1987

RX(1) OF 4     3 ***A***  ===>  B  +  C  +  ***D***...

                 :
                 :

RX(1)        RCT  3 A ***1129-62-0***
             PRO  B 118863-87-9, C 18108-38-8, D ***91-55-4***
             SOL  60-29-7 Et2O
             RGT  7719-12-2 PCl3
```

**Figure 7.** Example of a chemical reaction search using CASREACT.

**Table III.** Comparison of Reaction Information in CAS ONLINE and Beilstein

| feature | CA/REGISTRY | CASREACT | Beilstein |
|---|---|---|---|
| general features | | | |
| coverage | ≥1957 | ≥1985 | 1830–1979 |
| no. of substances | 10 million | | 3.4 million[a] |
| no. of reactions | | 750000 | 2.6 million preparations<br>0.4 million reactions |
| type of indexing | text, structure | RNs | text, structure |
| completeness | partly | yes | partly |
| quality | medium | good | medium |
| retrieval capabilities | | | |
| dictionary terms | yes | | yes |
| substructures | yes | | yes |
| role codes/subfields | | role codes | subfields |
| reaction sites | no | no | no |
| reaction conditions | no | only yield | yes |
| stereochemistry | text description | | only on DIALOG |

[a] The complete file contains 3.4 million substances covering the period from 1830 to 1979.

- to search for text terms (dictionary terms) or topology (connection tables) of specific substances or a class of substances
- to search for the preparation/chemical behavior of substances or classes of substances
- to search for the participants of a reaction
- to search for reaction parameter values (yield, conditions, etc.)
- to support the stereochemistry of the substances and of the reaction (in topological form)

- to upload and download structures and reactions
- to classify a substance as a starting material, solvent, catalyst, product, byproduct, or intermediate
- to search for single- and multistep reactions
- to search for successful or failed reactions
- to search for the reaction sites, i.e., transformation of reaction centers
- to protect (mask) functional groups in a reaction
- to search for similar structures and reactions

It is assumed that the storage and retrieval of text, numeric data, and connection tables (substructures) are available on online hosts. Reaction site searching, support of stereochemistry, multistep reaction searching, and retrieval of similar reaction are the major features which are still lacking, and they are discussed in more detail in the following subsections.

**Reaction Site Searching.** In CAS ONLINE (REGISTRY, CA, CASREACT, CJACS, etc.), chemical reactions can be searched on the basis of the connection tables of individual substances and the role which they play in the reaction process as products, starting materials, etc. This method of reaction searching may be called a substance-based reaction retrieval method. All reactions are searched and retrieved in which the specified substance(s) take(s) part. It is not possible, however, to search for the transformation of one functional group into another functional group, in particular it is impossible to search for reactions where a given functional group occurs only in the starting materials but not in the products. This feature may be called a reaction site search.[6,11-15,17,18,21,25] The difference between a substance-based reaction search and a reaction site search is illustrated in Figure 8. In the example, a reduction of an aromatic nitro group to an amino group is requested, and this should be the only change occurring in the reaction. A search based on substructure searches of the individual molecules retrieves all reactions where this reduction occurs (e.g., answers 1 and 2 in Figure 8) but also reactions with other changes in the molecular structure, e.g., the transformation of additional nitro groups (answer 3) or a diazo coupling (answer 3). The latter two reactions are not retrieved in a reaction site search. The results of such a reaction search are a subset of the answer set of the substance-based retrieval. In the case of a large database like CASREACT, it may happen that the substance-based reaction retrieval may yield many false hits, and the answer sets become very large due to this effect. Therefore, a reaction site search is an indispensable tool for the retrieval of chemical reactions. In contrast to online retrieval systems, both types of reaction retrieval are an intrinsic feature of the two in-house systems ORAC and REACCS.

A general reaction site searching capability should allow the user to search for both specific and generic reactions. It must be possible to search for generic substructures in a reaction as well as for generic reaction sites. Of course, such a system must be supported by a very comfortable graphical reactor editor.

**Support of Stereochemistry.** A large number of chemical substances require a stereochemical description. There are essentially two aspects of stereochemistry which should be supported by online retrieval systems.[25]

- distinction between stereoisomers
- preparation and reactions of stereoisomers

Stereoisomers are compounds which differ only by the orientation of the atoms in space. Some stereoisomers are mirror images of each other (enantiomers) and others are not (diasteromers). An online retrieval system should allow the user to refine his searches to the level of a specific stereochemical substance class, e.g., the derivatives of (+)-mannose. It should also be possible to choose between relative and absolute stereochemistry in a substructure search.

A synthetic chemist is also interested in the preparations or reactions of a given stereoisomer. This means that he has to find a reaction which yields, more or less exclusively, a specific stereoisomer. In some cases, this can be solved by searching for a substance with the respective stereochemistry. In other cases, however, it is necessary to search for stereoselective or stereospecific reactions. A stereoselective reaction is a transformation which yields predominantly one stereoisomer or one pair of enantiomers of several diastereomeric possibilities. An example for a stereoselective reaction is given in Figure 9 where the addition of bromine to 2-butene yields 2,3-dibromobutane. Both the product and the reactant can exist as diastereomers. The product can exist as a meso compound and a pair of enantiomers; the reactant can occur as a pair of geometric isomers. When the reaction starts from one diastereomer, e.g., the cis form, it results in the racemic form of the product, not the meso form. Further, if stereochemically different products give corresponding stereochemically different products, the reaction is called stereospecific. By this definition, the above reaction is also stereospecific, because the trans diastereomer reacts to the meso form (and not the racemate). The reaction shown in Figure 9 provides an example for such a reaction.

**Multistep Reaction Retrieval.** A chemical reaction may consist of many single steps, and it is clear that all individual steps of such a reaction must be recorded in a database. In the CASREACT database on STN, one may search for both singe- and multistep reactions which are recorded in a single document.[46] It is even possible to specify the number of steps or the maximum/minimum number which should be accepted in a reaction search. However, there are two limitations to this approach: (1) the reactions can only be searched based on the structures of the participating compounds and (2) all steps of the search reaction must occur in the same document, i.e., in a single literature reference. It is not possible to search for multistep chemical reactions where the different parts are recorded in different literature references. A reaction retrieval system should be able to retrieve multistep reactions independent of the literature references or the "documents" in the database. It must be possible to perform a reaction site search and retrieve all chemical reactions which satisfy the query transformation. In other words, there should be a mechanism that connects identical substances across all documents of the file. If the different parts of the reaction are recorded in different documents, then the retrieval software should be capable of locating all the steps which contribute to the reaction and finding the corresponding relationships. The use should be able to see both the overall reaction and the parts of the reaction together with their respective sources (references). This feature should be extended to work also between several different files (see Figure 10).

It is clear that it is not possible to store all possible reaction paths across documents. Besides the large storage requirement, it would also require a reload of the complete file with each update since many of the documents would be interconnected. A possible solution would be a reaction index directory which could include the connections between the reactions of the documents in the various files. However, this directory would be very difficult to maintain. In addition, there is also the problem of completeness: reactions which do not contain identical substances but only the derivatives of each other cannot be retrieved by using such a multifile reaction directory. In this case, a comparison of similar structures (similarity search) is also required.

**Similarity Searching of Reactions.** Although substructure and reaction site searching provide powerful tools for the online retrieval, they are not exactly identical with the views of the chemists. A concept which is more related to the thinking of

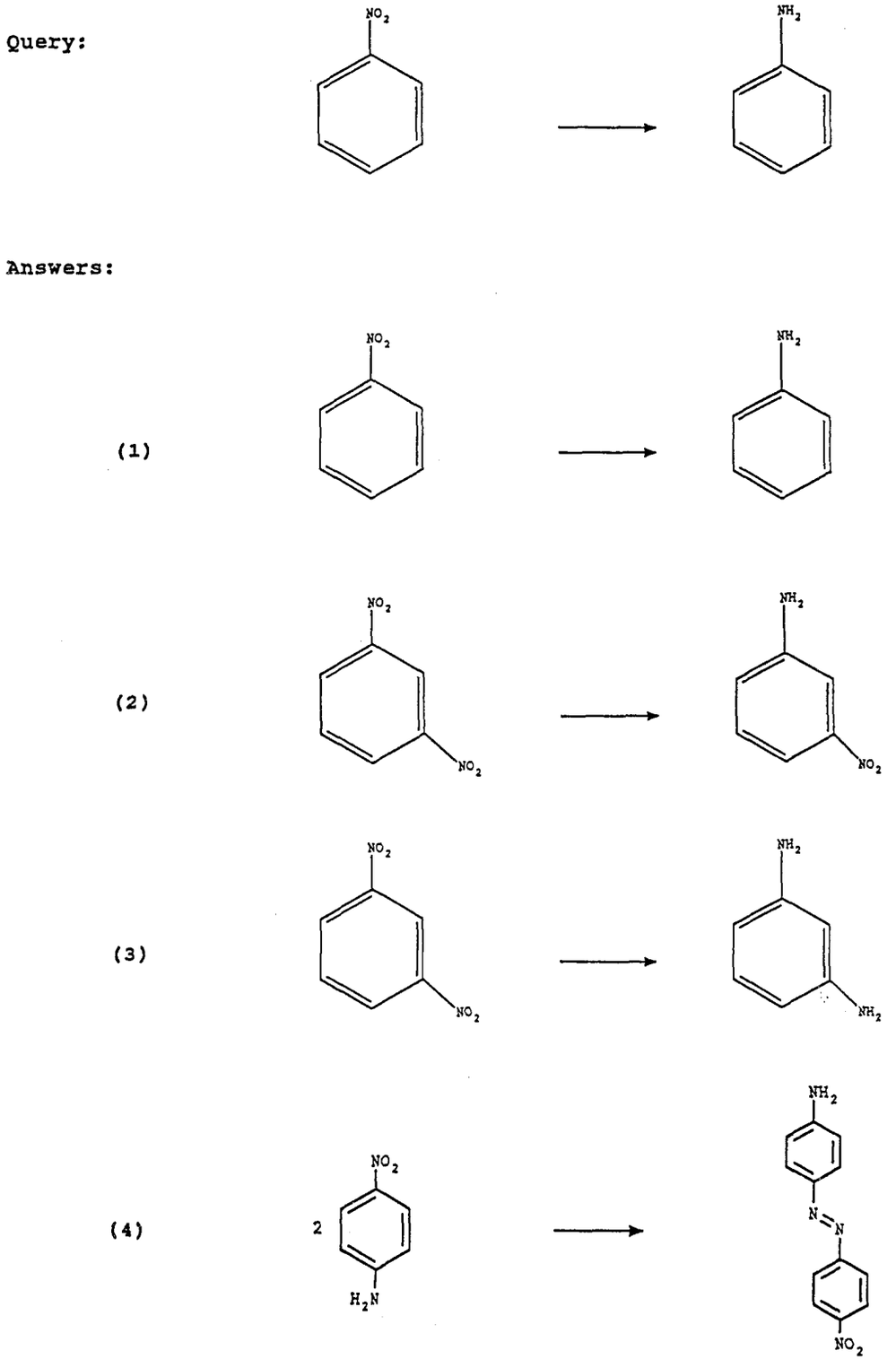



**Figure 8.** Illustration of reaction site searching.

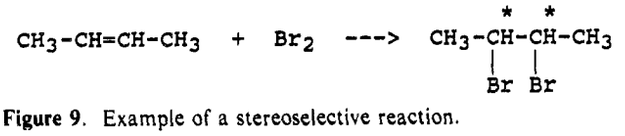



**Figure 9.** Example of a stereoselective reaction.

chemists is the idea of similarity.[47] Even though similarity of structures or reactions is not a concept which can be defined algorithmically, there are some aspects which could be handled by computer programs. Similarity of compounds comprises the following major classes of substances:

- Tautomers: This includes proton migration, mainly keto enol tautomerism, valence tautomerism, and charge mesomerism.
- Isomers: The most important groups are geometrical isomers and stereoisomers.
- Substances with similar molecular groups or atoms: This covers a broad variety of different aspects. It includes topological, geometrical, and functional similarity as well as isoelectronic groups and elements which belong to the same group of the periodic system.

A search for all tautomers of a given substance can be performed with substructure search systems if the starting materials are carefully defined. Isomers are also included in

| | | Reaction | | Doc. | File |
|---|---|---|---|---|---|
| Query Reaction: | A | ---> | X | | |
| Indexed Reaction: | A | ---> | B | 1 | I |
| | B | ---> | C | 1 | I |
| | C | ---> | D | 2 | I |
| | | : | | | |
| | L | ---> | M | 3 | II |
| | M | ---> | X | 4 | III |

**Figure 10.** Example of a multistep reaction indexing.

the answer sets of a substructure search. The problem today is the opposite—to distinguish between the stereoisomers. The last point comprises very different aspects. Substances with similar topological groups are covered by a substructure search. Geometrical similarity is probably more important in biochemistry where we deal with enzyme–coenzyme reactions. Several examples for the influence of geometrical similarity on conformational changes have been documented in the literature.[48-51] In this case, one is looking for geometrically similar substances which are accepted by the enzyme and influence the reaction either by inducing or by inhibiting a biochemical process. An example for the influence of geometrical similarity on the reactivity of an enzyme is the substance methotrexate vs tetrahydrofolic acid. The latter acts as a cofactor and is an important donator of C1 groups.[48] The substance methotrexate is geometrically similar to tetrahydrofolic acid and is also accepted by the enzyme, but it cannot transfer the C1 group, and thus acts as an inhibitor ($K_i \approx 10^{-10}$ M). This feature is currently not covered by any of the databases and retrieval systems since it requires a 3-D search capability.[52] However, it should be clear that geometrical similarity alone is not sufficient as a search tool.

Certain functional groups are reacting in a similar way, which means that they undergo the same reactions. However, this is dependent on the type of reaction and very difficult to implement in a retrieval system. Isoelectronic groups are groups containing the same number of valence electrons, e.g., $N_2$ and CO. These groups may show the same or opposite behavior in a chemical reaction. They can be searched and retrieved if they are defined by the user, i.e., the user has to specify a generic group which includes all groups with the same number of valence electrons. An automatic inclusion of these groups is not possible today, but it does not represent a principal problem for the information retrieval systems. The similar behavior of elements which belong to the same group in the periodic system can also be taken into account by the definition of a generic group through the user.

## SUMMARY

There are several databases available through online hosts which contain reaction information. Among the most important ones are the CA, REGISTRY, CASREACT, and BEILSTEIN Files. Only CASREACT is a true reaction database providing the most comprehensive search capabilities for reactions. The other databases, however, are important sources of information, especially if the searcher is interested in substance preparation. The retrieval possibilities of the online hosts with respect to reactions (transformations) are still rather basic. In this paper, the most important requirements for a reaction data service have been outlined. Support of stereochemistry is a general feature which is also required for structure coding and substructure searching. Reaction site searching, multistep searching, and the support of similarity in reaction retrieval are major features to be developed for online hosts. Since there are several large reaction databases under development, it can be assumed that the future retrieval capabilities of online hosts have to account for new features as described in this paper.

## ACKNOWLEDGMENT

The funding of this work by the Federal German Ministry for Research and Technology is greatly acknowledged.

## REFERENCES AND NOTES

(1) Fugmann, R.; Bitterlich, W. Reaction documentation using the GREMAS system. *Chem.-Ztg.* **1972**, *96* (6), 323–30. Fugmann, R.; Kusemann, G.; Winter, J. H. The supply of information on chemical reactions in the IDC system. *Inf. Process. Manage.* **1979**, *15*, 303–23.
(2) Schier, O.; Nuebling, W.; Steidle, W.; Valls, J. System for the documentation of chemical reactions. *Angew. Chem., Int. Ed. Engl.* **1970**, *9* (8), 599–604.
(3) Ziegler, H. J. Documentation of organic reactions. *Ind. Chim. Belge* **1967**, *32*, 88–92; A new documentative system for organic reactions. *Ind. Chim. Belge* **1968**, *33* (9), 744–9; Organic reactions: a new technique of organic reaction documentation. *Inf. Chim.* **1968**, *No. 41*, 22–4, 27–8, 31–4.
(4) Armitage, J. E.; Crowe, J. E.; Evans, P. N.; Lynch, M. F.; McGuirk, J. A. Documentation of chemical reactions by computer analysis of structural changes. *J. Chem. Doc.* **1967**, *7* (4), 209–15.
(5) Harrison, J. M.; Lynch, M. F. Computer analysis of chemical reactions for storage and retrieval. *J. Chem. Soc. C* **1970**, *15*, 2082–7.
(6) Valls, J.; Schier, J. Chemical Reaction Indexing. In *Chemical Information Systems*; Ash, J. E., Hyde, E. E., Eds.; Horwood Ltd.: Chichester, 1975.
(7) *Communications, Storage and Retrieval of Chemical Information.* Ash, J. E., Chubb, P. A., Ward, S. E., Welford, S. M., Willett, P., Eds.; Ellis Horwood Ltd.: Chichester, 1985.
(8) Meyer, E. Information Science in Relation to the Chemist's Needs. In *Chemical Information Systems*; Ash, J. E., Hyde, E., Eds.; Ellis Horwood: Chichester, 1975.
(9) Willett, P. Computer techniques for the indexing of chemical reaction information. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 156–8.
(10) Willett, P. The evaluation of an automated indexed, machine-readable chemical reactions file. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 93–6.
(11) McGregor, J. J.; Willett, P. Use of maximum common subgraph algorithm in the automatic identification of ostensible bond changes occurring in chemical reactions. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 137–40.
(12) Boother, J. A graphical approach to reaction retrieval in organic chemistry. *Chem. Br.* **1985**, *21*, 68–9.
(13) Barnard, M. J. M. Problems and solutions for retrieval of reaction information. *Colloq. Inf. Chim. (C.R), 2nd*, **1986**, 200–3.
(14) Willet, P., Ed. *Modern Approaches to Chemical Reaction Searching.* Blackmore Press: Shaftesbury, 1986.
(15) Deroulede, A. An update on computer-based systems providing information on chemical reactions and syntheses. *Inf. Chim.* **1987**, *289*, 143–6.
(16) Funatsu, K.; Endo, T.; Kotera, N.; Sasaki, S.-I. Automatic recognition of reaction site in organic chemical reaction. *Tetrahedron Comput. Methodol.* **1988**, *1* (1), 53–69.
(17) Johnson, A. P. Reaction indexing: An overview of current approaches. In *Chemical Structures: The Universal Language of Chemistry*, Proceedings of the Chemical Structures Association Conference, Nordwijkerhout, The Netherlands, June 1987; Springer-Verlag: Berlin, 1988.
(18) Behnke, C.; Bargon, J. Computer-assisted topological analysis and completion of chemical reactions. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 228–37.
(19) REACCS (REaction ACCess System) is a chemical reaction database management system from Molecular Design Ltd. for in-house files.
(20) French, S. E. Our reaction access system. *CHEMTECH* **1987**, *17*, 106–11.
(21) Moock, T. E.; Nourse, J. G.; Grier, D.; Hounshell, W. D. The implementation of atom–atom mapping and related features in the Reaction Access System (REACCS). In *Chemical Structures: The Universal Language of Chemistry*, Proceedings of the Chemical Structures Association Conference, Nordwijkerhout, The Netherlands, June 1987; Springer-Verlag: Berlin, 1988.
(22) Grethe, G.; del Rey, D.; Jacobson, J. G.; VanDuyne, M. Reaction indexing in an integrated environment. In *Chemical Structures: The Universal Language of Chemistry*, Proceedings of the Chemical Structures Association Conference, Nordwijkerhout, The Netherlands, June 1987; Springer-Verlag: Berlin, 1988.
(23) Borkent, J. H.; Oukes, F.; Nordik, J. H. Chemical reaction searching in REACCS, SYNLIB, and ORAC. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 148–50.
(24) Zass, E.; Mueller, S. New possibilities for research on organic chemical reactions: comparison of "in-house" databank systems, REACCS, SYNLIB, and ORAC. Chimia **1986**, *40*, 38–50.
(25) Kasparek, S. V. Computer graphics and chemical structures: database management systems, CAS Registry, Chembase, REACCS, MACCS-II, Chemtalk. John Wiley: New York, 1990.
(26) ORAC (Organic Reactions Accessed by Computer) is a chemical reaction database management system from ORAC Ltd. for in-house files.
(27) CAS ONLINE is a set of databases grouped around the Registry File and the CA File of literature abstracts. In addition, it comprises

full-text files like CJACS (Chemical Journals of the American Chemical Society) and CASREACT, a file of chemical reactions.
(28) Dittmar, P. G.; Farmer, N. A.; Fisanick, W.; Haines, R. C.; Mockus, J. The CAS ONLINE search system: 1. General system design and selection, generation, and use of search screens. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 93–102.
(29) Schulz, H. From CA to CAS ONLINE: The data collection of Chemical Abstracts Service and their use. VCH: Weinheim, 1985.
(30) Vander Stouw, G. G. Potential enhancements to the CAS Chemical Registry system. In *Chemical Structures: The Universal Language of Chemistry*, Proceedings of the Chemical Structures Association Conference, Nordwijkerhout, The Netherlands, June 1987; Springer-Verlag: Berlin, 1988.
(31) Farmer, N.; Amoss, J. Farel, W.; Fehribach, J.; Zeidner, C. The evolution of the CAS parallel structure searching architecture. In *Chemical Structures: The Universal Language of Chemistry*, Proceedings of the Chemical Structures Association Conference, Nordwijkerhout, The Netherlands, June 1987; Springer-Verlag: Berlin, 1988.
(32) The Beilstein database contains the structures and factual data of all well-defined organic substances described in Beilstein's *Handbook of Organic Chemistry* or from the (unchecked) literature excerpts of the recent primary literature.
(33) *The Beilstein Online Database*; Heller, S., Ed.; ACS Symposium Series; American Chemical Society: Washington, DC, 1990; in press.
(34) Heller, S. R. A Survey of Reaction Databases. *Proceedings of the 13th International Online Information Meeting*, London; Learned Information: Oxford, 1989.
(35) The Chemical Reaction Documentation Service (CRDS) produced by Derwent Ltd. and offered through Maxwell Online is based on Theilheimer's *Synthetic Methods of Organic Chemistry* and the *Journal of Synthetic Methods*.
(36) Finch, A. F. The Chemical Reactions Documentation Service: ten years on. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 17–22 and references therein.
(37) Dana, R. C. Where Do They All Come From? Appropriate Coverage of the Literature For A Chemical Reactions Database. *Proceedings of the 13th International Online Information Meeting*, London; Learned Information: Oxford, 1989.
(38) Jochum, C., Beilstein Institute, private communication, 1990.
(39) The ZIC (VEB Zentrale Informationsverarbeitung Chemie) is a file of 1.6 million preparations covering the years 1980–1988 with 250 000 updates per year. It includes the topological reaction transformation; the corresponding literature references; and a few factual data, like melting point, boiling point, and yield. (P. Löw, CHEMODATA, private communication, 1990.)
(40) Gasteiger, J.; Weiske, C. An Integrated Information System on Chemical Reactions. *Proceedings of the 13th International Online Information Meeting*, London; Learned Information: Oxford, 1989.
(41) Parlow, A.; Weiske, C.; Gasteiger, J. *ChemInform*—an integrated information system on chemical reactions. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 400–402.
(42) Beach, A. J.; Dabek, H. F.; Hosansky, N. L. Chemical reactions information retrieval from Chemical Abstracts Service publications and services. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 149–55.
(43) Ai, C.; Blower, P. E.; Ledwith, R. H. Extraction of chemical reaction information from primary journal text. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 163–9 and references therein.
(44) Blake, J. E.; Dana, R. C. CASREACT: More than a million reactions. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 394–399.
(45) North, S. Chemical reaction information: what are the user needs? *Colloq. Inf. Chim. (C.R.), 2nd*, **1986**, 204–11.
(46) Blower, P. E.; Chapman, S. W.; Dana, R. C.; Erisman, H. J.; Hartzler, D. E. Machine generation of multi-step reactions in a document from single-step input reactions. In *Chemical Structures: The Universal Language of Chemistry*, Proceedings of the Chemical Structures Association Conference, Nordwijkerhout, The Netherlands, June 1987; Springer-Verlag: Berlin, 1988; and references therein.
(47) Moock, T. E.; Grier, D. L.; Hounshell, W. D.; Grethe, G.; Cronin, K.; Nourse, J. G.; Theodosiou, J. Similarity searching in the organic reaction domain. *Tetrahedron Comput. Methodol.* **1988**, *1* (2), 117–28 and references therein.
(48) Stryer, L. *Biochemistry*; W. H. Freeman: New York, 1988.
(49) Beardsley, G. P.; Moroson, B. A.; Taylor, E. C.; Moran, R. G. A new folate antimetabolite, 5,10-dideaza-5,6,7,8-tetrahydrofolate is a potent inhibitor of de novo purine synthesis. *J. Biol. Chem.* **1989**, *264*, 328–33.
(50) Taylor, E. C.; Harrington, P. M. A convergent synthesis of 5,10-dideaza-5,6,7,8-tetrahydrofolic acid and 5,10-dideaza-5,6,7,8-tetrahydrohomofolic acid. An effective principle for carbonyl group activation. *J. Org. Chem.* **1990**, *55* (10), 3222–7.
(51) Biellmann, J. F. Chemistry and structure of alcohol dehydrogenase: some general considerations on binding mode variability. *Acc. Chem. Res.* **1986**, *19* (10), 321–8.
(52) Murrall, N. W.; Davies, E. K. Conformational freedom in 3-D databases. 1. Techniques. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 312–6.