CASREACT: More than a Million Reactions

JAMES E. BLAKE and ROBERT C. DANA*

Chemical Abstracts Service, P.O. Box 3012, Columbus, Ohio 43210

Received July 20, 1990

CASREACT, an online database providing access to chemical reactions reported in the journal literature, was introduced by Chemical Abstracts Service in 1988. The file aims at broad, rapid coverage of reactions coupled with specialized search software that enables users to focus their searches. More than one million reactions from over 65 000 documents are accessible through CASREACT. Interaction with the CAS Registry File and graphic displays aid both searching and interpretation of CASREACT information.

This paper begins with a discussion of reaction information in the CA File and the origins of CASREACT. It then focuses on CASREACT as it exists today in terms of content and coverage, search features, the various display options for viewing the reaction information, and other general capabilities of the file. It also discusses user feedback with regard to CASREACT and actions taken or being planned as a result of that feedback.

REACTION INFORMATION IN THE CA FILE

The reaction information in the CA File falls into three categories: (1) terms from the title, abstracts, and keywords; (2) compounds indexed and the text associated with them; and (3) controlled vocabulary index terms, particularly reaction terms. Title, abstract, and/or keyword terms always reflect the terminology used by the author; keywords may also include standardized terms. The use of controlled vocabulary terms is based on the author's emphasis shown in the title, abstract, and full text of the document.

Compounds are indexed by their CAS Registry Number (RN) with either a simple text 'prepn. of' or additional words showing preparation, properties, and/or uses for a product. This information is combined with a number of synonyms for 'prepared' (in context) to algorithmicly associate the searchable suffix 'P' with CAS Registry Numbers (e.g., 1234-56-7P). This pointer makes the retrieval of prepared compounds straightforward in the CA File. CAS has indexed starting materials routinely since 1976. Reactants are associated with 'reaction of' and other reaction-type terms. When a compound is both prepared and further reacted, the text shows both ideas, e.g., 'preparation and reaction of, with ...'. This type of text allows the proper assignment of the 'P' and shows the further reaction.

Generally, only reactants and products are indexed for the CA File. Author emphasis occasionally will cause reagents and (more rarely) solvents to be indexed as well. Catalysts which are prepared are always indexed, but the common catalysts are indexed only as a result of author emphasis. Text terms with any of these reaction participants will make their role obvious.

Controlled vocabulary terms—general subject terms in printed *Chemical Abstracts* (CA)—are hierarchically assigned for CA, with the most specific term being chosen for input. Thus, named reaction terms, if they have been selected as controlled terms, are indexed in preference to more generic terms, e.g., *Diels-Alder cycloaddition* is preferred to *Cycloaddition reaction*. When the novelty of the paper is in the catalyst(s) used, the controlled term will involve the catalyst, e.g., *Cycloaddition reaction catalysts* would be input instead of *Cycloaddition reaction*. None of these terms are associated with the specific compounds in the document.

The reaction information in the CA File comes from all journals, patents, and reports covered in printed CA. Overall, the CA File has been and remains an excellent source of reaction documents.

ORIGINS OF CASREACT

The service now known as CASREACT resulted from repeated user requests for substructure-based reaction-retrieval capabilities. Users were aware that reaction information was available in the CA File as described above, but they wanted more. Specifically, they wanted

- precise linkage of reactants, products, and other reaction participants
- information about formation or cleavage of specific bonds
- the ability to search for substructural transformations
- graphic display of reactions

CAS activities in this area began as a research project in the late 1970s. Dr. P. E. Blower, Jr., has over the years led a variety of research efforts at CAS that resulted in many of the capabilities of CASREACT. Initial efforts focused on the ability to capture reaction information as part of the analysis of documents, to retrieve registry structures and create from them reaction site information, to link input of single-step reactions from a given document into the multistep protocols found in the original papers, and finally to develop a search system to allow the user to retrieve this information.

COVERAGE AND CONTENT

CASREACT has taken a broad, inclusive approach to database building primarily because the adequate provision of multistep syntheses, not single reactions, requires a far broader selection of reactions from a document. Unlike CA, CASREACT has been built from a core list of journals abstracted in the 14 organic sections of CA, whose titles are

21. General Organic Chemistry
22. Physical Organic Chemistry
23. Aliphatic Compounds
24. Alicyclic Compounds
25. Benzene, Its Derivatives, and Condensed Benzenoid Compounds
26. Biomolecules and Their Synthetic Analogs
27. Heterocyclic Compounds (One Hetero Atom)
28. Heterocyclic Compounds (More than One Hetero Atom)
29. Organometallic and Organometalloidal Compounds
30. Terpenes and Terpenoids
31. Alkaloids
32. Steroids
33. Carbohydrates
34. Amino Acids, Peptides, and Proteins

File building began in late 1984 so that any document from the core list published after January 1, 1985, would be in-

0095-2338/90/1630-0394$02.50/0

cluded. There are no current plans to include material earlier than 1985. The initial core list contained 102 journals, but new journals and journals added as a result of user and analyst suggestions have increased the total to the present 111 journals. Changes in journal titles have been carefully tracked to assure continuing coverage. The original selection rules for CASREACT were so broad that almost all candidate documents from these journals were included. The only documents not included were those that contained no reactions. "Novelty" has not been a primary selection criterion.

Feedback from potential users together with a better understanding of the database caused CAS to adjust some of the selection criteria in mid-1986. All coverage changes have been intended to focus CASREACT on useful synthetic chemistry. For example, reactions studied for theoretical reasons are now excluded from coverage. Products formed in very low yields are generally excluded.

CASREACT is still a highly inclusive database because almost all reactions from documents selected for coverage are included in the file. However, the focus on synthetic utility has suggested that

- Well-known chemistry such as simple salt formation, preparation of hydrates or solvates, esterification, or the preparation of characterizing derivatives need not be included. Such reactions will be included to keep intact a long synthetic path.
- Labeling reactions will be included when there is a novel method.
- Procedures that are automated, usually Merrifield preparations, will not be included.
- Reactions that are optimized will be covered under either the optimum conditions or conditions stressed by the author.
- Novel preparations from a mechanistic or spectral study will be included rather than the reaction being studied.

The coverage and content decisions about selection of documents and the inclusion of reactions have, since late 1984, led to the following characteristics for CASREACT:

- All reaction participants for a selected reaction are included, even if the compound(s) is(are) not indexed for CA.
- More than 70% of candidate documents get CASREACT coverage.
- The average CASREACT document contains 15 single-step reactions.
- Over 90% of CASREACT documents include reaction participants not covered in the CA File or printed indexes.
- During 1989, more than 15 000 documents with over 207 000 single-step reactions were added to the database.
- Users will find reactions no matter how long the synthetic path from the desired reactant to the desired product.
- Searches based on general transformations or generic substructures are likely to retrieve a large answer set.

Recent surveys of users worldwide have suggested to us that further refinements of CASREACT coverage are appropriate. Hence, proposed changes for the selection of documents and reactions for CASREACT when fully implemented will include new, modified, or improved reactions from the full range of documents covered in CA. Multistep sequences will continue to be provided when such reactions are included in the sequence. Steps in the sequence that are individually of minor synthetic significance will usually be combined into a single transformation. Document analysts will be relying more heavily on author indications of novelty. For example, a searchable 'key step' term will be available in the Note field to highlight steps in a multistep sequence that the author indicates is crucial to the sequence. Conversely, author emphasis on a step which would otherwise be combined will ensure single-step treatment.

The outcome of these refinements is expected to be:

- coverage of all new, modified, or improved reactions in the documents covered by CA
- compacted multistep sequences
- fewer single-step reactions added per year

CAS will continue to fine-tune the coverage of CASREACT, both in terms of the reactions included and the areas of chemistry covered. As always, the expressed needs of our users will guide future changes.

SEARCHING IN CASREACT

CASREACT reaction searching may actually begin in any one of three CAS ONLINE files. First, searches may be entered directly in CASREACT; second, users may begin in the Registry File; or third, they may begin in the CA File.

Starting Directly in CASREACT. Users may search specific CAS Registry Numbers or the accession number for a given document. The substance search in CASREACT may be qualified by the role of the reaction participants. The roles are Reactant, Reagent, Product, Catalyst, and Solvent. There are also two combination roles: Reactant/Reagent and Anything-but-Product. For example:

```
=> SEARCH 50-00-0/RCT
=> SEARCH 112:20534/AN
```

Reagents. Although reagents have been indexed for CASREACT since 1984, user feedback showed a need to clarify the distinction between reactants and reagents. Hence, CAS has recently begun to use the following distinction:

- Compounds contributing one or more carbons to the product are reactants.
- Compounds which contribute no carbons are reagents.

There are no plans to review older reactions to reclassify compounds according to this definition. However, in September 1989, a new combination role 'RRT' was created which allows the user to search for a compound as either a reactant or reagent with a single search qualifier. The use of the 'RRT' role should remove any concern that the user might have about whether a substance is a reactant or a reagent.

Solvents. The Solvent role (SOL) may be assigned to a compound even if another role is also assigned. Thus, in an esterification it is common that the esterifying alcohol is also the solvent. In such cases, the alcohol will be given both the reactant and solvent roles.

Catalysts. The Catalyst role (CAT) may be assigned based on author usage even though another role would be more appropriate. An example would be common Lewis acids, which often are present in quantities that would generally cause them to be considered 'reagents', but are called 'catalysts' by the authors.

In addition to individual substances in CASREACT having the Catalyst role, the ANY catalyst feature allows the chemist to discover any catalysts which have been used for a specified reaction. The term 'ANY/CAT' linked to the reaction search will find reactions which are catalyzed. The display of these reactions will provide a line formula, an acronym, or the CA systematic name for the catalyst. This allows the user to see specific catalysts which can provide the desired structural change without having to guess what possible compounds might be good catalysts.

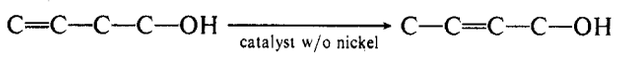
Starting Reaction Searches in the Registry File. Reaction searches for some or all of the reaction participants in Registry may be approached in either or both of two ways: structure searches or name/formula searches. The searches create answer sets corresponding to reaction participants. These

answer sets are then used in the actual reaction searches in CASREACT.

To illustrate, here is an example of a reaction search which poses the question, "Are there catalysts which are not nickel-containing compounds that might be used to isomerize 3-butenols to 2-butenols?" Catalysts often are difficult to find by substructure searching. Indeed, one of the most powerful search methods in the Registry File is to search in the Molecular Formula Index for any compounds containing one or more atoms of the desired metal. These answer sets can be large and will usually contain many inactive metal salts, oxides, etc., as well as useful catalysts. Refining the answer set can be a major challenge.

Two structure searches and a text search are run in the Registry File in preparation for the reaction search in CASREACT. The structure searches each include two screens to limit the retrieval; screen 2082 limits a structure search to substances indexed in CASREACT; screen 1836 indicates only primary alcohols may be retrieved. The text search finds all of the nickel-containing substances in CASREACT (CASREACT/LC). After completing these Registry File searches, CASREACT is called, and the reaction search is entered using the answer sets created by the searches in the Registry File connected by (L) operators. The (L) operator indicates that all participants are included in the same reaction sequence. The (NOTL) is used to preclude the nickel compounds from the answers.

Question:



In the Registry File:

```
      SEARCH Reactant structure C=C—C—C—OH
L7    225 answers
      SEARCH Product structure C—C=C—C—OH
L9    542 answers
      Both queries include the following screens:
      Screen 2082 CASREACT
                  1836 Primary alcohols
      SEARCH NI>=1/MF and CASREACT/LC
L10   2472 answers
```

In CASREACT:

```
      SEARCH L7/RRT (L) L9/PRO (L) ANY/CAT
      (NOTL) L10/CAT
L11   2 answers
```

The user should first display answers using the DISPLAY SCAN feature which shows the first hit reaction and the title from a randomly selected hit document. This free display is used to verify the search results. Having verified that the search is retrieving correct results, the searcher can then display the retrieved answers more fully. The several display options are discussed later.

Starting Reaction Searches in the CA File. These searches use usual CA File search techniques to retrieve documents containing information pertinent to the reaction being sought. Once a satisfactory answer set has been created, then the search is narrowed to those documents which are also in CASREACT. At this point the searcher may enter CASREACT and search the CA File answer set from CA and display the reactions within the documents.

```
=> FILE CA
=> SEARCH DIELS ALDER REACTION #/IA
L1   8491 DIELS ALDER REACTION #/IA
=> SEARCH L1 AND CASREACT/OS
L2   1842 L1 AND CASREACT/OS
=> SEA L2 AND 1990/PY
L3      4 L2 AND 1990/PY
=> FILE CASREACT
...
=> SEA L3
L4      4 L3
```

The above example retrieves four CASREACT documents published in 1990 and indexed in CA with 'Diels-Alder reaction'.

DISPLAYS IN THE CASREACT FILE

Once a search has been completed, CASREACT offers a variety of ways to examine those answers. Each reaction display is made up of three parts: the map (1), the diagram (2), and the summary (3). Users may display the map and diagram only or the map and summary only. (See Figure 1.)

The map contains generic identifiers for the reactants and products and also specifies the reaction number within the document. It also tells the user which single-step reactions make up any multistep reaction being displayed.

The diagram shows the structures of the reactants and products. The diagrams may be displayed either on a graphics terminal with Plot10 graphics capability or on a text terminal.

The summary contains the roles and CAS Registry Numbers for all of the reaction participants. It also includes any text information about the reaction, e.g., safety information. Chemical names or representations are given for the catalysts, solvents, and reagents.

The hit substances are highlighted in both the map and the summary. The document information (CA File bibliographic, abstract, and index data) may also be displayed with the reaction information.

The default display shows both the first hit reaction and the bibliographic information for the document. There are about 20 other predefined display formats that allow the user to display as much or as little reaction data as desired. A new STN feature, SET FORMAT, also makes it possible for users to define their own display format. The user-defined display format can even be designated as the default display format for CASREACT by using the command SET DEFORMAT.

In the following example, L4 is the answer set number from a reaction search. A user display format will be defined (.MYOWN) (1) and designated as the default display (2) in CASREACT on a permanent basis. This user format consists of the first hit reaction format (FSAM), the bibliography format (CBIB), and the abstract format (AB). After designating the default format, a display of an answer is requested (3). (See Figure 2.)

```
(1) => SET FORMAT .MYOWN FSAM CBIB AB
    SET COMMAND COMPLETED
(2) => SET DEFORMAT .MYOWN PERMANENT
    SET COMMAND COMPLETED
(3) => DISPLAY L4 2
```

CASREACT users may display all of the "hit" reactions (HIT format) within a document or all of the reactions in the document (RX format). Any specific reaction may be displayed in full [RX(6) format for the 6th reaction in the document].

Multistep reactions are displayed (in the diagram) as if they were single-step reactions. If desired, users may also display the diagrams of all of the steps [RXL(42) format for the multistep reaction 42]. (See Figure 3). Users may also display specific single steps of the multistep reaction [RX(5) format for the step that is the 5th reaction in the document].

Many multistep reactions are very complex with several reactions converging at various points in the sequence. The displays of these sequences are also complex. The phrase COMPOSED OF LINEAR SEQUENCE...AND LINEAR SEQUENCE... at the beginning of the display indicates the presence of a complex multistep reaction. The specific reactions involved in each chain of the sequence are indicated in

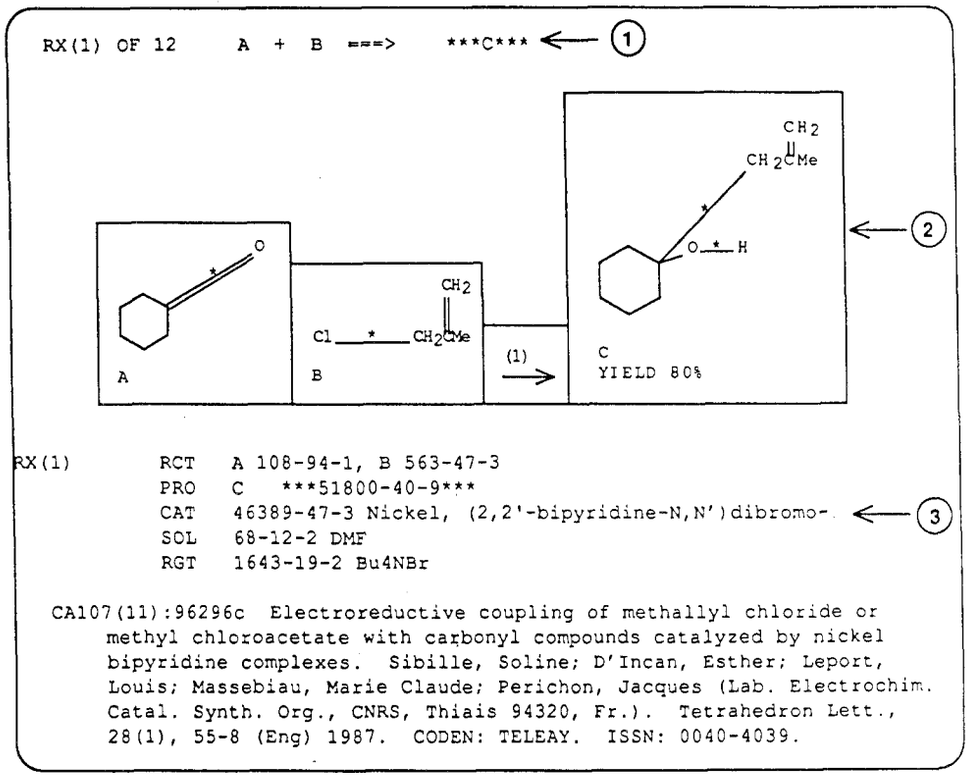


Figure 1.

the message in place of the periods shown above.

Often a document will include several hit reactions. The OCC display format reports the reaction numbers of the hit reactions and the number of hit reactions in the shortest reaction path. It is good practice before using the HIT or RX formats to use the OCC format to determine the number of reactions these displays may involve. The OCC format also reports the number of PATH and SPATH terms in each reaction and the number of PATH and SPATH reactions. The SPATH format shows the map and diagrams of all reactions in the shortest path, and is particularly useful when the answer is a complex multistep sequence. The PATH formats work best when the query contains both a reactant and a product. (See Figure 4.)

FAILED REACTIONS

Since the start of CASREACT in 1984, document analysts have been instructed to include reactions that did not work. There are two criteria for inclusion:

1. The author must specify the expected product of the reaction.
2. The author must show interest in the failure of the reaction.

Generally, these reactions are steps in a reaction sequence which unexpectedly failed. The authors usually find alternative methods and speculate on, or even investigate, the cause or failure. Such failures show limits previously not known to particular reactions or structural transformations. As such, they are important adjuncts to a successful synthetic methodology. Such failures are rarely input to the CA File, but they are easily found in CASREACT by searching 'failed reaction' as a text search. The HIT reactions will all be failed reactions and will be displayed with a slash (/) through the reaction arrow.

YIELD INFORMATION

CASREACT did not contain yield information initially because of the following characteristics of yields in the literature:

- Yields are not given for all compounds.
- Few yields in the journal literature are optimized.
- Authors may report crude yields, yields calculated from spectra, or yields of isolated and purified products in the same paper.

Discussions with users made it clear that yield is a valued criterion. So, CAS began to include yield information in CASREACT in late 1986. Yield data is range searchable in the YD field and displays in the diagram. The fact that not all papers include yield data and that almost 2 years of the file had no yield information required that users be given a mechanism to include answers to a given structural query whether or not yield data was present. Users who wish to include answers of this type include NONE/YDT as part of their query.

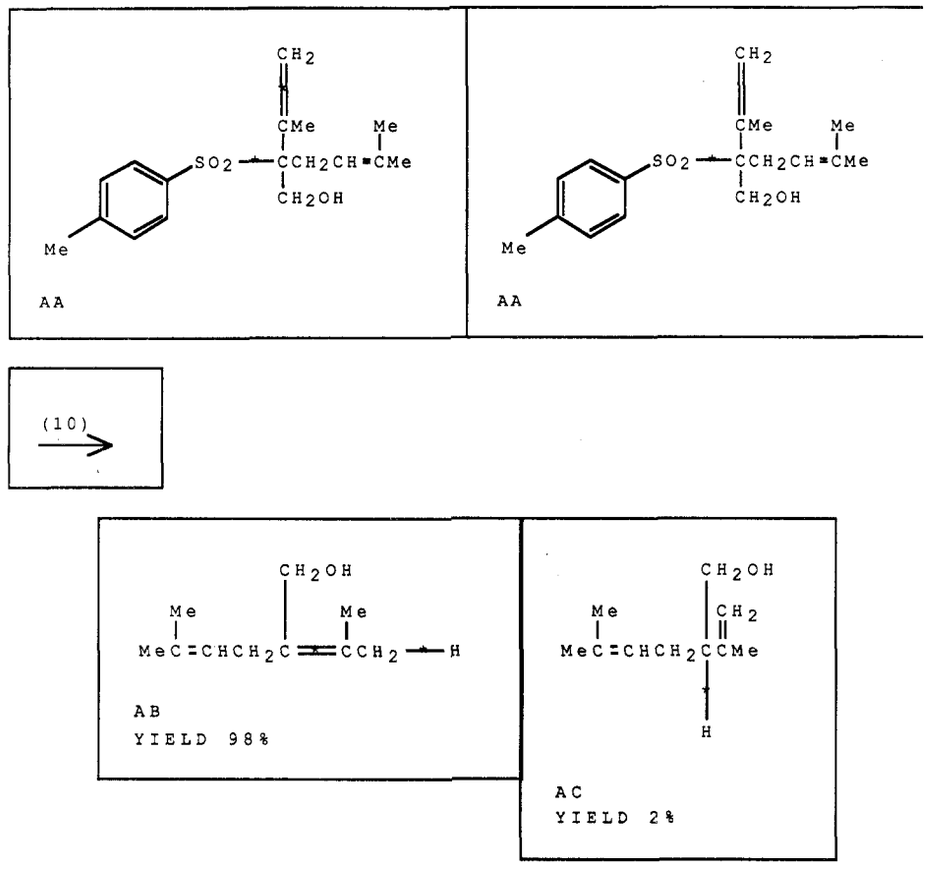
The start of yield input pointed out additional problems. First, many reactions give mixtures for which the author has given only an overall yield, not a yield for each product. When that is the case, CASREACT analysts put the yield information in the Note field. Unlike the /YD, this information is not range searchable. The identifying word 'overall' is placed next to the percent yield for the product mixture in the Note. Second, mixtures of stereoisomers, regioisomers, or optical isomers may have a yield reported for the mixture and further information which indicates the ratio of compounds in the mixture, e.g., a 60:40 mix of two stereoisomers. In such cases, the yield for the mixture is searchable, but the ratio is not; the ratio is posted as display information with the yield. Thus, if one got a 70% yield of *E* and *Z* isomers which were found to be 2:1 *E*:*Z*, the searchable yield would be 70, with (66) and (33) posted with the yield for the appropriate isomer.

SAFETY INFORMATION

In CASREACT, safety information is present in the Note field of the reaction and is searchable. The term 'safety' is used. Other terms, some of which are used in conjunction with the term 'safety', in CASREACT include forms of 'exploded...

L4 ANSWER 2 OF 2
COPYRIGHT (C) 1990 AMERICAN CHEMICAL SOCIETY

RX(10) OF 37 ...2 ***AA*** ===> ***AB*** + AC



RX(10) RCT 2 AA ***113767-89-8***
PRO AB ***3304-27-6*** , AC 1845-51-8
CAT ***79500-51-9*** Palladium, [[2-[bis(4-methylphenyl)phosphino]ethyl]diphenylphosphine-

P,P']dichloro-, (SP-4-3)-, ***789-25-3*** Silane, triphenyl-
SOL 109-99-9 THF
RGT 22560-16-3 Borate(1-), triethylhydro-, lithium, (T-4)-
NTE 20.degree., 3 min

CA108(17):150723a Regio- and stereoselective synthesis of allylic and homoallylic alcohols by the reductive desulfonylation of allylic sulfone derivatives. Application to the syntheses of (.+-.)-lavandulol and isolavandulol. Inomata, Katsuhiko; Igarashi, Susumu; Mohri, Mitsunobu; Yamamoto, Taku; Kinoshita, Hideki; Kotake, Hiroshi (Fac. Sci., Kanazawa Univ., Kanazawa 920, Japan). Chem. Lett. (4), 707-10 (Eng) 1987. CODEN: CMLTAG. ISSN: 0366-7022.

AB Regio- and stereoselective desulfonylation of RCH:CHCH(SO2C6H4Me-p)CHR1OH [R = Me, PhCH2, Me2CH; R1 = PhCH2CH2, Me(CH2)7, Me, Me2CH] provided a convenient method for the prepn. of RCH2CH:CHCHR1OH and RCH:CHCH2CHR1OH. The title compds. were prepd. in excellent yields by this method.

Figure 2.

or explos...', 'toxic', 'pyrophoric', and 'carcinogen...'.

REACTION CONDITIONS

Some reaction condition information is available for searching in CASREACT. Although specific values for such reaction conditions as time, temperature, pressure, or pH are not indexed, the document analysts have often input terms such as 'photochem...', 'gas phase', or 'electrochem...' to describe reaction conditions which are generally not obvious from the reactants or products. The phases which are most commonly found in the Note field are

anaerobic	in the dark
biochem.	in vacuo
buffered soln.	inverse addn.
electrochem.	photochem.
enzymic	radiochem.
equil	solid state
fermn.	thermal
gas phase	ultrasound
high pressure	

Such terms should be primarily used to refine queries which have retrieved large answer sets. These terms are input as the result of some author emphasis and hence are not necessarily present for all reactions.

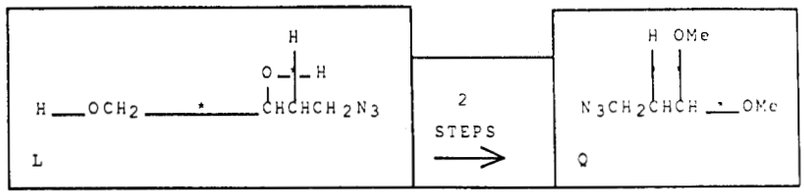
SUMMARY

CASREACT has already recorded its millionth reaction. New coverage guidelines will extend CASREACT as a reactions database drawn broadly from the literature covered by CA. CASREACT provides the user precise retrieval of

```
=> D RX(42)

L1  ANSWER 1 OF 5


RX(42) OF 159 COMPOSED OF RX(5), RX(6)
L  ===>  Q...
```




```
RX(5)     RCT   L 114642-93-2
          PRO   O 114642-94-3
          SOL   7732-18-5 Water
          RGT   7790-28-5 NaIO4

RX(6)     RCT   O 114642-94-3
          PRO   Q 114642-95-4
          CAT   ***7664-93-9***   H2SO4
          SOL   67-56-1 MeOH
          RGT   ***67-56-1***   MeOH, 7758-98-7 CuSO4
```

Figure 3.

```
=> D OCC

L1  ANSWER 1 OF 5


NUMBER OF HIT REACTIONS            61
NUMBER OF REACTIONS IN PATH         4
NUMBER OF REACTIONS IN SPATH        5
FIELD                           COUNT
RX(6)                               2
RX(12)                              2
RX(13)                              2
RX(42)                              2
RX(43)                              2
RX(48)                              2
RX(49)                              2
                .
                .
                .
RX(157)                             2
RX(158)                             3
RX(159)                             3
NUMBER OF HIT REACTIONS            61
NUMBER OF REACTIONS IN PATH         4
NUMBER OF REACTIONS IN SPATH        5
```

Figure 4.

reaction information by access points that are not available in the CA File.

User feedback is being sought and used to improve both the database content and retrieval capabilities. Reaction-site searching is the next major enhancement planned for CASREACT, with more enhancements expected as the unique relationship between CASREACT, the Registry File, and the CA File is improved. More refined techniques to help users ask questions and find answers in their pursuit of chemical reaction information are coming in the future.