# INFORMATION RETRIEVAL: SCIFINDER® AND SCIFINDER® SCHOLAR

**Damon D. Ridley**

*University of Sydney, Australia*

WILEY

John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester,
West Sussex PO19 8SQ, England

Telephone (+44) 1243 779777

Email (for orders and customer service enquiries): cs-books@wiley.co.uk
Visit our Home Page on www.wileyeurope.com or www.wiley.com

Reprinted with corrections April 2003



***Other Wiley Editorial Offices***

John Wiley & Sons Inc., 111 River Street, Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 33 Park Road, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 22 Worcester Road, Etobicoke, Ontario, Canada M9W 1L1

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

***British Library Cataloguing in Publication Data***

A catalogue record for this book is available from the British Library

ISBN 0-470-84350-0 (Hardback)

ISBN 0-470-84351-9 (Paperback)



Typeset in 10/12pt Times New Roman by Laserwords Private Limited, Chennai, India
Printed and bound in Great Britain by Antony Rowe Ltd, Chippenham, Wiltshire
This book is printed on acid-free paper responsibly manufactured from sustainable forestry in which at least two trees are planted for each one used for paper production.

# Contents

# Preface

Since the early 1990s, the Chemical Abstracts Service (CAS) has been developing an intuitive 'point and click' interface for scientific information. Subsequently, SciFinder® *http://www.cas.org/SCIFINDER/SCIFINDER* was released in 1995 as desktop software for commercial organizations. The academic version, SciFinder® Scholar *http://www.cas.org/SCIFINDER/SCHOLAR* became available in 1998.

The products have different usage conditions and pricing structures, but these are of little concern in this text. Instead, this text is intended to help scientists get the maximum benefit from SciFinder and SciFinder Scholar *through an understanding of the content and functionality*. The products are developing along parallel (but not identical) lines, so at times they will contain some differences. However, in the vast majority of cases, they function in the same way, so this text is equally a guide for users of SciFinder as well as SciFinder Scholar.

SciFinder Scholar contains all the functionality present in SciFinder, and so this text mainly refers to the former product, and to the academic environment in which it is used. Additional functionality that occurs in SciFinder is indicated by reference only to SciFinder. So when SciFinder appears in the text, academic users need to know that the functions discussed presently apply only to the commercial version.

Since SciFinder Scholar is an intuitive end-user interface, the obvious question is 'Why a book on Information Retrieval: SciFinder Scholar?' Primarily, the answer is that SciFinder Scholar is an *interface* and that behind it lies the large and complex scientific literature. This text is about the

understanding of the literature and of how SciFinder Scholar may be used as a *creative* interface.

Actually, perhaps an equally valid question is 'Why *not* a book on information retrieval?' since while students in university courses spend around 5000 hours in lectures, tutorials, and laboratory sessions learning the rigors of scientific method and countless technical facts, they are presented with very little rigorous analysis of the scientific literature. In a way, the World Wide Web and its search engines reinforce the unreasonable expectation that quality answers may be obtained without knowledge of the literature. However, how reliable and how comprehensive is data from Web search engines, and what important research may have been missed?

Everything taught in college courses ultimately comes from the scientific literature that in turn contains so much information that it cannot all be covered. So why is not scientific information retrieval an integral part of all college courses? And why is it that so many staff and students who are so committed to the understanding of, and curious about, science may be reluctant to invest even a few hours in the development of a better appreciation of the scientific literature and of how to search and use it effectively?

The answers to these questions may be in part historical, cultural, and financial. The answers may be in part through the size and nature of science itself, because without doubt a basic appreciation of science and scientific method is required before the complex issues of the scientific literature are presented. Nevertheless, at some time, scientists have to address issues relating to information retrieval, and different scientists may have different needs. The information needs of a postgraduate student who is learning about research and the research field may be quite different from those of the senior professor, or the industrial scientist.

Effective information retrieval requires considerable skills and is an intellectual challenge. In obviating the need for detailed knowledge of search mechanics, SciFinder Scholar takes the tedious work out of information retrieval and allows the scientist to focus on an understanding of the content of the literature. It allows the scientist to think creatively about the difficult part of searching, that is, knowing how to ask the right questions in the first place.

SciFinder Scholar is very different from all other information retrieval tools. The most user-friendly alternative search tools, including Web search engines, still require some knowledge of truncation, of proximity searching, and of synonyms and so forth. They then interpret the question literally, and rarely do they offer any guidance on how to proceed. Depending on what the scientist really wanted, they may or may not produce comprehensive and/or



precise answers. SciFinder Scholar works differently in that it automatically handles, *inter alia*, truncation, proximity searching, and synonyms. To the experienced 'power' searcher, this may initially be considered a limitation and indeed in some cases the search algorithm may not precisely execute the user's intentions. However, the sophisticated search commands that are required to narrow down answer sets to manageable and affordable proportions in on-line systems with transactional pricing structures are not needed in SciFinder Scholar. Now answers may be displayed in full at no extra cost and may be refined manually according to the user's needs.

SciFinder Scholar also works differently in that it *guides the searcher*. Indeed, when a question is initially asked, SciFinder Scholar does not give a straight answer! Instead, it guides the searcher by producing a set of options. It tells the user 'go down this option and you will find a specified number of records, whereas this other option provides a different number of records.' In other words, the user chooses a path on the basis of the actual number of hits, but the choice is not irrevocable and the user may always return to narrower or broader answer sets.

Once a path has been chosen, SciFinder Scholar has *creative analyze options*. For example, SciFinder Scholar will provide histograms of the different index terms, or document types, or author names, or publication years, and the scientist chooses refinement paths accordingly. In chemical structure searching, SciFinder Scholar automatically interprets the query to allow for different structure conventions and representations. The amount of prior knowledge needed indeed is minimal! If answers are too numerous, then SciFinder Scholar again guides the searcher through analysis tools.

Because it is so different, users with some experience in other systems may initially think that SciFinder Scholar is limited. In fact, the reverse is true and this text explains why. For example, while the user may want to search for articles on the inhibition of HIV replication in humans by Erik de Clercq from 1990 onward, SciFinder Scholar actually would not allow it in one initial search. It must be taken in steps. Even an experienced on-line professional, using sophisticated search techniques, may take some time to obtain all the correct hits on this topic, while a relatively inexperienced user of SciFinder Scholar may do almost as well in a few minutes.

This text is not designed to explain the mechanics of searching or of data processing. These are detailed in the SciFinder Scholar manual, which may be downloaded from the Web *http://www.cas.org/SCIFINDER/SCHOLAR2002/sfsguide.pdf*, and in the numerous help messages available through the help icons at the top of the computer screen. Help on the use of SciFinder Scholar is also available from other sources, some of which are given in Appendix 1.

'Why a book on SciFinder Scholar?' Indeed! SciFinder Scholar is so much fun and is so creative. It opens up the world's scientific literature in a unique way, and this text is designed to help the user gain maximum benefit from this unique product.

This text starts with an outline of the basic content of the databases and the way SciFinder Scholar searches these databases. It explains why certain answers are retrieved, and how features of SciFinder Scholar may be used to refine or broaden searches in a predictable way. It explores different options to the solutions of problems, and above all it encourages scientists to be creative and to think carefully about how to approach problems. SciFinder Scholar is a *research tool* and not just a search tool!

A theme throughout this text is to encourage scientists to apply 'scientific method' to their information retrieval. To achieve this, scientists should understand the background of their 'experiments' (initial searches), and so Chapter 1 summarizes some basic aspects of the scientific literature and then describes aspects of the content of the databases searched. Luckily, SciFinder Scholar handles many issues of content automatically, but in some cases, a brief understanding of content will help in asking the initial query and with the processing of answers. Chapters 2 and 3 work through the two principal search options, namely, searching for topics and for substances, respectively. They explain why it is preferable to search initially at broad levels and then to use, *inter alia*, the very valuable analyze/refine options to focus on key answers.

Searches may also be started from a number of other perspectives, and there are many ways to work through answers. These are considered in Chapter 4. Chapter 5 considers some advanced search strategies particularly in the chemical, medical, and biological sciences, while Chapter 6, on searching for chemical reaction information, is to help the synthetic chemist obtain information in the complex area of chemical reactions. Finally, the text contains many appendixes, which have extensive information of interest in specialized areas.

All chapters contain exercises that reinforce the most important teaching points. Answers to these exercises are available through the website *http://www.wiley.co.uk/ridley*. There often is no single correct way to search but at least the answers give insights into some possibilities. This website contains information on updates and text for Appendices 8, which thus may be printed to any size required. Please also refer to this website for a description on how text for this Reprinted Edition was updated to accommodate the October 2002 release of SciFinder Scholar.



The SciFinder concept is highly innovative ... and brilliant! The implementation of the concept took years to achieve and is a result of the combined efforts of the CAS staff and scientists worldwide. Particular thanks are due to Bob Massie and all his staff at CAS for making it all possible. I thank also the University of Sydney, which generously has allowed me the time to travel extensively and to study the science of information retrieval. In recent years, particular thanks are due to my friends and colleagues Les Field and Sev Sternhell. Finally, over the years many have shared their knowledge of information retrieval with me and have directly or indirectly contributed to this text. Accordingly, I thank Brian Cannan, Barry Dunne, Thomas Haubenreich, Carolynn Larsen, John Martin, Janice Mears, Ken Ostrum, Paul Peters, Kirk Schwall, Eric Shively, Joe Uffner, Mike Walsh, Fred Winer, Kurt Zielenbach, and above all Eva, Lloyd, Andrew, Nicholas, and Natasha.

*Damon D. Ridley*
*October 2001*

Chapter 1

# What is SciFinder® Scholar?

## 1.1 Introduction

Knowing how to search and use information is a key factor in successful teaching and research. However, the information world is so complex that staff and students are often overwhelmed. Further, quality information may be expensive and academic institutions may find it hard to provide even a basic level of service.

It has not always been like this. Perhaps, until the third quarter of the twentieth century, universities were able to give access to the major journals, indexing tools, and reference materials needed by their scholars. Much has changed with the onset of the electronic age, with declining budgets, with increases in the price and the number of journals, and with the development of new research areas.

One excellent solution is offered through SciFinder Scholar. Before understanding how to use SciFinder Scholar, it is necessary to understand what Scholar is and how it fits into the world of information retrieval.

## 1.2 The Primary and Secondary Literature

### *1.2.1 Primary Literature: Journals*

The number of journals in the sciences alone presents a major problem. For example, one of the world's major science database producers, the Chemical Abstracts Service (CAS), covers around 8000 journals. For original articles with the publication year 1999, the CAS bibliographic database (CAPLUS) contains almost 900 000 records. Almost two-thirds of these articles are

from journals produced by just over 3000 publishers, and approximately one-third of these records were published by scientific societies, academies, institutes, associations, and various university presses. This vast literature clearly presents problems not only for scientists who need to keep up with relevant literature in their field but also for librarians whose function is to serve their scientific community.

Much has been written on the subject of scientific publication, on the pressures on journal subscriptions, and on the influence of governments, capital markets, and technology. However, it should be remembered that scientists have benefited from both the society and the commercial publishers and are indeed fortunate to have such excellent primary and secondary sources of scientific information that are the envy of other disciplines (particularly the humanities and social sciences). Scientists' problems are not in the area of collection or indexing but rather in the size and cost of basic information needs. It helps if scientists and librarians work cooperatively with publishers, who are now responding much more favorably to the needs of institutions. For example, libraries are finding that substantial discounts are available through package deals with the major publishers.

While further changes in the scientific information industry will occur, what will always underpin the industry will be the need for the scientists to publish their discoveries and in turn for scientists to search for literature in their area.

Increasingly, publications will appear less in print and more directly through the Internet. Among the advantages of the latter is that it not only allows more direct access at the desktop but also allows linking of documents, for example, through citations. This worldwide web of publications will grow, and as the web search engines become more sophisticated, scientists will be able to search the vast primary literature more directly and conveniently.

### *1.2.2 Secondary Literature: Abstracting and Indexing Services*

Abstracting and indexing services will continue to provide valuable services not only because they provide access to literature that predates the Web but also because they will continue to build databases that will be vital to the information needs of the scientists.

There are many databases that cover the sciences, and Table 1.1 summarizes some bibliographic databases that currently contain an excess of two million records. Whether the scientist should search these databases, or even more subject-specific and smaller databases, essentially depends on the type of research involved and to the cost and the availability of the service. If there



**Table 1.1** Major scientific bibliographic databases and database producers. SciFinder Scholar contains CAPLUS and MEDLINE

| Database producer | Database name[a] | Area covered | Number of records[b] |
|---|---|---|---|
| American Institute of Aeronautics & Astronautics | AEROSPACE | Aerospace research | 2.2 |
| National Agricultural Library | AGRICOLA | Agriculture | 3.6 |
| BIOSIS | BIOSIS® | Life sciences | *12.7 |
| CAB International | CABA | Agriculture and related sciences | 3.9 |
| CAS | CAPLUS | Multidisciplinary science | 18[c] |
| Engineering Information Inc. | COMPENDEX | Engineering and technology | 4.8 |
| Elsevier Science | EMBASE | Medical and biological sciences | 8.4 |
| US Department of Energy | ENERGY | Energy | 3.9 |
| American Geological Institute | GEOREF | Geosciences | 2.3 |
| International Atomic Energy Agency | INIS | Nuclear research and technology | 2 |
| Institution of Electrical Engineers | INSPEC | Physics, electronics, and computing | 6.6 |
| US National Library of Medicine | MEDLINE | Medical sciences | 11.4 |
| Institute for Scientific Information | SCISEARCH® | Multidisciplinary science | 18.8 |
| US National Library of Medicine | TOXLINE® | Toxicology of drugs and chemicals | 2.7 |
| Derwent Information Ltd | WPINDEX | Patents | 10 |

[a] On the STN network.
[b] Approximately (million) to end of 2000.
[c] However, a major backfile was added in 2001, and the current figure (September 2001) is 19.7 million.

is a single, inexpensive, and specific database that completely covers the information needs of the searcher, then the searcher indeed is fortunate.

However, for comprehensive retrieval, the scientist usually needs either to search a number of databases or to search databases that cover broad areas. The former option is most readily realized either through on-line or localized networks, and among the world's major international on-line

networks in the sciences are Dialog *http://www.dialog.com*, Questel/Orbit *http://www.questel.orbit.com/english/index.htm*, and STN *http://www.cas.org/stn.html*. The effective use of on-line networks requires very substantial training, and access to these networks generally is through a pay-as-you-use cost structure, which is not necessarily ideal for institutions that need to provide access to information to a large number of users. Nevertheless, it helps if scientists are aware of the additional opportunities provided through these sources and work with their library staff when access through them may be of specific benefit.

In the sciences, four of the world's biggest databases (i.e. those with more than 10 million records in Table 1.1) are produced by BIOSIS *http://www.biosis.org/home_deluxe.html*, CAS *http://www.cas.org*, the National Library for Medicine (NLM) *http://www.nlm.nih.gov*, and the Institute for Scientific Information (ISI) *http://www.isinet.com*. Generally, scientists looking for broad coverage explore some of these databases first. It is important for scientists to know the areas that the different products cover, their specific features, and their various benefits if they are to perform efficient and effective searches.

Searching for information requires many skills and the process involves the application of the 'scientific method' in a manner similar to the conduct of a laboratory experiment. The analogy is summarized in Figure 1.1, and the suggestion is that university courses should include instructions on all

| Step | Laboratory Experiment | Information retrieval |
|---|---|---|
| 1 | Knowledge of the subject | Knowledge of the coverage and content of databases, of potential alternative information resources, and of search options |
| 2 | Requirement for new knowledge | Requirement for information on a topic |
| 3 | Proposal for research or experiment | Conceptualization of initial search query |
| 4 | Conduct of experiment | Initial search |
| 5 | Observation and interpretation of results | Careful examination of initial answer sets |
| 6 | Revision or improvement of experimental design | Revision of search query based on observation of initial answers |
| 7 | Revised experiment | Exploration of alternative search options |
| 8 | Final outcome of experiment | Final answer set |

**Figure 1.1** The parallel between use of scientific method in the laboratory and in information retrieval.



aspects of information retrieval and on the various products available, and instruction on the technical aspects of the scientific discipline! This will help scientists follow the 'scientific method' in information retrieval.

## 1.3 The SciFinder Scholar Option

SciFinder Scholar *http://www.cas.org/SCIFINDER/SCHOLAR* provides desktop access to several databases from CAS and to the bibliographic database from the US NLM (Table 1.2), and thus covers information from two of the world's largest database producers. The only other single source of all of these databases is the STN network.

The databases have very extensive coverage across the sciences. To illustrate this, Figure 1.2 shows a breakdown of records in CAPLUS (>30 000 records) and MEDLINE® (>14 000 records) for Schools and Departments in

**Table 1.2** Databases in SciFinder Scholar

| Database | Database name[a] | Coverage | Number of records[b] |
|---|---|---|---|
| CAS bibliographic database | CAPLUS | More than 9000 journals, and patents from more than 30 organizations from 1907[c] | >22 million |
| NLM bibliographic database | MEDLINE | More than 4300 journals from 1958[d] | >13 million |
| CAS chemical substance database | REGISTRY | All substances from 1967[c] | >47 million (substances) |
| CAS regulated chemicals database | CHEMLIST | Substances on national inventories and some US State lists[e] | >220 000 |
| CAS chemical catalogs database | CHEMCATS | Substances in chemical catalogs from more than 600 suppliers[f] | >5.6 million (substances) |
| CAS chemical reaction database | CASREACT | Mainly organic reactions from selected journals from 1907[g] | >6 million reactions |

[a] With the exception of the MEDLINE file, all files are available on the STN network only.
[b] February 2003.
[c] Although many pre-1967 substances are also covered *http://www.cas.org/casdb.html*.
[d] *http://www.nlm.nih.gov/hinfo.html*.
[e] *http://www.cas.org/CASFILES/chemlist.html*.
[f] *http://www.cas.org/CASFILES/chemcats.html*.
[g] *http://www.cas.org/CASFILES/casreact.html*.

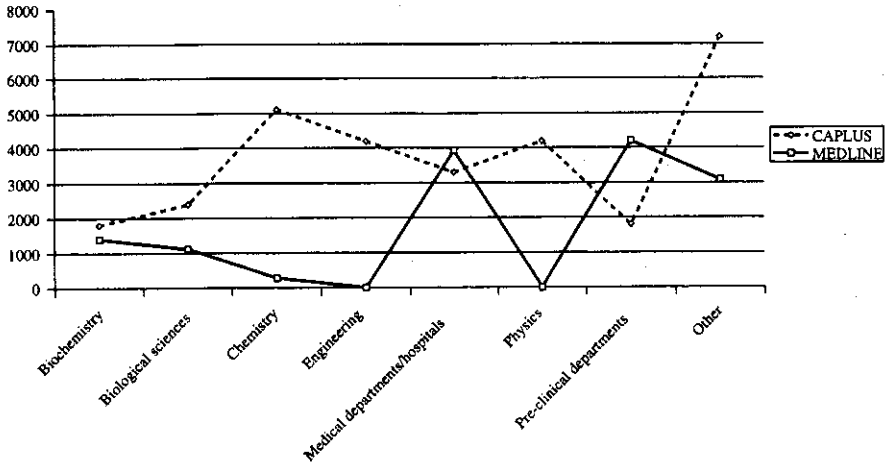



**Figure 1.2** Number of records for discipline areas in Australian universities in CAPLUS (>30 000 records) and MEDLINE (>14 000 records) for 1995 to 1999. Data was obtained through searches in the Corporate Source Field in the CAPLUS and MEDLINE files on STN. Records from individual schools and departments were then combined into the broad categories illustrated. (Note: 'Other' in CAPLUS includes *inter alia* Environmental Sciences, Geosciences, Material Sciences, and entries in which only the University was listed. 'Other' in MEDLINE includes *inter alia* Dentistry, Health Sciences, Nursing, Psychology, and Veterinary Science.)

Australian universities from 1995 to 1999. While the breakdown may vary in different countries, nevertheless, it is clear from Figure 1.2 that CAPLUS covers a much broader range of disciplines than that is of interest just to chemistry schools, and that in combination, these two databases cover nearly all disciplines in the sciences very extensively!

Another aspect of coverage is that CAPLUS currently contains data for more than three million records from patents. For the publication year 2000, almost one-third of the records in the database are from patents and more than 40% of the new organic substances registered in the CAS substance database are from patent documents. No search is complete without paying attention to the very extensive patent literature!

In addition to breadth of coverage, SciFinder Scholar has many other features including currency (the major CAS databases are updated daily and MEDLINE is updated four times a week), links to full text patents and to electronic journals, options to links to in-house collections, unique search options including structure and reaction searches, and a search interface that guides the searcher through possibilities. With respect to the secondary



literature, SciFinder Scholar currently provides solutions to most of the information needs of the staff and students in the sciences in a unique way, and certainly in a way not matched by any other single desktop information tool.

## 1.4 Types of Databases in SciFinder Scholar

SciFinder Scholar is an interface that requires minimum training, and it certainly obviates the need for detailed knowledge of search mechanics and search commands. However, some knowledge of the databases is valuable. First, it is helpful to understand that two of the databases (CAPLUS and MEDLINE) are bibliographic databases, three (REGISTRY, CHEMLIST, and CHEMCATS) are substance databases, and the last (CASREACT) is a chemical reaction database.

Some entries in bibliographic databases are text from authors and some are text from indexers. Consequently, *it helps if the searcher first thinks how an author may have described the research in the title and abstract.* In practice, there is little consistency among authors, and Chapter 2 discusses how SciFinder Scholar manages this issue primarily through automatic application of synonyms and truncation and searching for terms either within the same sentence or within the whole record.

It also helps if the searcher considers *how an indexer may have applied the index policies* of the database producer to enter text into the record. In this chapter, basic indexing principles are described, although scientists should continue to learn about indexing by careful examination of actual records. Chapter 2 discusses how SciFinder Scholar offers searchers a number of unique options to take advantage of indexing through the analysis of records by index term, supplementary term, section codes, CAS Registry Numbers, CAS Roles, and MEDLINE headings, and allowable qualifiers.

## 1.5 CAS Bibliographic Database (CAPLUS)

Up-to-date information on the content of CAPLUS is available at *http://www.cas.org/casdb.html*. In broad terms, the database contains more than 22 million records for original publications from a variety of primary sources. While the major source is journal articles (74%), there is an extensive coverage of patents (16%) and conferences (5%); among the remaining document types are dissertations, reports, and books.

Records from about 8000 journals are included. Generally, only records from those original articles that contain information related to 'chemistry'

are fully indexed, but 'chemistry' is *very broadly defined* by CAS, and the outcome is partly illustrated in Figure 1.2! Further information on the coverage is available through the 'more than chemistry' brochures, which may be obtained at *http://www.cas.org/SCIFINDER/SCHOLAR*.

Since 1995, CAS has included entries for all articles in approximately 1350 core journals (*http://www.cas.org/sent.html*), and details of the patents

**Bibliographic Information**

**Spectroscopy and photosensitization of sapphyrins in solutions and biological membranes.** Roitman, Leonid; Ehrenberg, Benjamin; Nitzan, Yeshayahu; Kral, Vladimir; Sessler, Jonathan L.. Dep. Phys., Bar Ilan Univ., Ramat-Gan, Israel. Photochem. Photobiol. (1994), 60(5), 421-6. CODEN: PHCBAP ISSN: 0031-8655. Journal written in English. CAN 122:26806 AN 1995:202539 CAPLUS

**Abstract**
A spectroscopic and photophys. study of three new sapphyrin mols. is presented. The sapphyrin backbone that was derivatized to make them water sol. possesses an absorption band around 700 nm, a desired property for biol. photosensitization. We studied the absorption and fluorescence spectra, from which evidence for aggregation in solvents of different polarities was obtained. The extent of aggregation is correlated with the nature of the attached moiety. The abs. quantum yields of singlet oxygen prodn. were measured, with 1,3-diphenylisobenzofuran as a model target, and were 0.13-0.18 in ethanol. The binding consts. to liposomes and to cells were detd. spectroscopically and were found to correspond to the hydrophobicities of the compds., with an addnl. effect, ascribed to the sugar moiety, which was found in the case of one of the sapphyrins. The efficiency of photodamage to Staphylococcus aureus by sapphyrins and hematoporphyrin was equiv., on the basis of cells killed per μg of sensitizer in the incubation mixt.

**Indexing** -- Section 8-3 (Radiation Biochemistry)
Cell membrane
Laser radiation
Membrane, biological
Photodynamic action
Photosensitizers
Staphylococcus aureus
(sapphyrins spectroscopy and photosensitization in solns. and biol. membranes)

Porphyrins
Role: BAC (Biological activity or effector, except adverse); THU (Therapeutic use); BIOL (Biological study); USES (Uses)
(sapphyrins, sapphyrins spectroscopy and photosensitization in solns. and biol. membranes)

158012-18-1
Role: BAC (Biological activity or effector, except adverse); FMU (Formation, unclassified); BIOL (Biological study); FORM (Formation, nonpreparative)
(sapphyrins spectroscopy and photosensitization in solns. and biol. membranes)

159858-86-3
159858-87-4
Role: BAC (Biological activity or effector, except adverse); THU (Therapeutic use); BIOL (Biological study); USES (Uses)
(sapphyrins spectroscopy and photosensitization in solns. and biol. membranes)

7782-44-7, Oxygen, biological studies
Role: BAC (Biological activity or effector, except adverse); THU (Therapeutic use); BIOL (Biological study); USES (Uses)
(singlet; sapphyrins spectroscopy and photosensitization in solns. and biol. membranes)

**Supplementary Terms**
sapphyrin photosensitization biol membrane

**Figure 1.3** Sample record from CAPLUS. Copyright the American Chemical Society and reproduced with permission.



covered are available at *http://www.cas.org/EO/caspat.html*. While no single database contains all the world's scientific journal and patent literature, the users of SciFinder Scholar may be very confident that in CAPLUS, they are searching one of the world's most comprehensive sources of scientific information!

An example of the content of a single record is shown in Figure 1.3. After the title and bibliographic information, the records contain the abstract, followed by the indexing. Titles and abstracts are usually those written by authors, although in some cases (most notably for patents) the titles and abstracts may be rewritten by indexers to better reflect the technical aspects of the original article.

One of the greatest benefits of searching databases is that users may take advantage of the indexing added by the database producer. The advantages of indexing are many, including *inter alia*:

- *systematic terminology* so that the different terms used by authors to cover a particular topic may effectively be searched under a consistent term for that topic. In this way, use of a few index terms will produce more comprehensive answer sets;
- *precise entries into topics*, since often the database producer will add the index term only if it relates to a focus of the original research; and
- *additional entries to the record*, for example, not only does the systematic indexing of chemical substances enable precision searches to be accomplished but also important substances mentioned in the article will be entered in the indexing even though no mention of them may be made in the title or abstract written by the author.

Most database producers also arrange their indexing in a hierarchy, so that searchers may easily go from one index term to terms that more broadly, or more precisely, describe the topic. The CAS hierarchy of terms is included in the CA Lexicon®, and part of the lexicon for the *Staphylococcus* genus is shown in Figure 1.4. At this stage, it helps if SciFinder Scholar users note that in some on-line systems, searches for the *Staphylococcus* genus may be conducted more generally through successive levels from Bacillaceae, to gram-positive bacteria, to Bacteria. Alternatively, the searcher may search more specifically through individual species such as *Staphylococcus aureus* and *Staphylococcus capitis*. However, searching with the aid of a thesaurus is quite specialized and although at present SciFinder Scholar does not automatically search hierarchies, it does have a number of ways to take advantage of index terms.

```
=> E STAPHYLOCOCCUS/CT+ALL
E1        11126   BT4  Bacteria (Eubacteria)/CT
E2         1043     BT3  Gram-positive bacteria (Firmicutes)/CT
E3            0       BT2  Low G+C gram-positive bacteria/CT
E4           66         BT1  Bacillaceae/CT
E5         6115           -->  Staphylococcus/CT
                  HN  Valid heading during volume 66 (1967)to present.
E6                       UF   Staphylococci/CT
E7                       UF   Staphylococcosis/CT
E8           77          NT1  Coagulase-negative Staphylococcus/CT
E9            7          NT1  Peptococcus saccharolyticus/CT
E10           0          NT1  Staphylococcus afermentans/CT
E11           4          NT1  Staphylococcus arlettae/CT
E12           0          NT1  Staphylococcus auranticus/CT
E13       16322          NT1  Staphylococcus aureus/CT
E14           1            NT2  Staphylococcus aureus altilis/CT
E15           7            NT2  Staphylococcus aureus anaerobius/CT
E16           6            NT2  Staphylococcus aureus aureus/CT
E17           1            NT2  Staphylococcus aureus gallinae/CT
E18           3            NT2  Staphylococcus aureus hemolyticus/CT
E19          17          NT1  Staphylococcus auricularis/CT
E20          75          NT1  Staphylococcus capitis/CT
E21           3            NT2  Staphylococcus capitis capitis/CT
E22           3            NT2  Staphylococcus capitis ureolyticus/CT
E23          18          NT1  Staphylococcus caprae/CT
E24         215          NT1  Staphylococcus carnosus/CT
E25           1            NT2  Staphylococcus carnosus utilis/CT
```

**Figure 1.4** Example of the CA Lexicon in the CAPLUS file on STN for the index heading *Staphylococcus*. The lexicon gives broader terms (BT), and narrower terms (NT) in the hierarchy, and the number of records in the database for the specific index terms. (Note: only the first 25 of the 80 hierarchical terms are shown here.) Copyright the American Chemical Society and reproduced with permission.

CAS indexing appears in three main parts. First the database is divided into 80 sections (*http://www.cas.org/PRINTED/sects.html*) and each record is given the section and subsection number that is considered to be most relevant to the overall content of the article. For the record in Figure 1.3, the Section Code is 8-3 (Radiation Biochemistry). These broad section codes may be useful for the refinement of answer sets, and the analysis and refinement by section codes are discussed in Chapter 2, Section 2.5.4.

The main CAS indexing, which immediately follows the listing of the section codes, consists of index headings and CAS Registry Numbers, and each of these has text-modifying phrases (which are included in parentheses).

Although a full list of indexing terms is available and the information processionals need to understand all the levels of indexing to achieve optimal searches through on-line networks, it helps if SciFinder Scholar users understand only some basic principles.



### 1.5.1 Index Terms: Subjects and Animal, Plant, and Microorganism Species

More than 25 000 subject index terms for subjects are used, and relevant terms are entered in records in which a major aspect of the original article relates to the CAS index policy for this term. In Figure 1.3, the five subject index terms are cell membrane; laser radiation; membrane, biological; photodynamic action; and photosensitizers. The CAS also systematically enters names of species when the original article describes something new about the species. More than 120 000 species' names are used, and, for example, the species' index term listed in Figure 1.3 is *Staphylococcus aureus*.

Analysis of answers by index terms is easily achieved in SciFinder Scholar and offers another important option for the refinement of answers (Chapter 2, Section 2.5.4). The searcher need not know the indexing in advance, as Scholar displays the indexing in a graphical format once initial answers have been obtained. Remember that index terms are systematic entries that may be used to help obtain more precise answers or to help with identifying terms to enter into search queries; so, part of the 'scientific experiment' involved with searching for information requires careful scrutiny of indexing.

### 1.5.2 Index Terms: Substance Class Headings

The CAS has defined many substance class headings and these are entered when the original article describes something new about the substance class or about a number of specific substances that are described by this class. For example, if reactions of a variety of individual alkenes are mentioned in the original text, the substance class heading 'Alkenes' will be used. (In Figure 1.3, the substance class heading 'Porphyrins' is used.)

Substance class headings always have *additional* levels of indexing through CAS Roles (see Section 1.5.4), and in Figure 1.3, the CAS Roles, namely, BAC, THU, BIOL, and USES are added to the heading 'Porphyrins' to indicate more precisely the aspect of research on porphyrins that is included in the original article.

### 1.5.3 Index Terms: Substances

The CAS Registry Numbers underpin the CAS Registry System, and the only complete source of CAS Registry Numbers is the REGISTRY database (Section 1.7), which may be searched in SciFinder Scholar in a number of ways (Chapter 3). Each unique substance is given a unique CAS Registry

Number and when something new is reported on the substance in the original article, the CAS Registry Number is added as an index term in the database record. Not every substance in the original article is indexed and indeed some key CAS indexing policies relating to substances are indicated in Table 1.3.

**Table 1.3** Some indexing policies for substances in CAPLUS

| Indexing policy | Example | Implications |
|---|---|---|
| The CAS Registry Number is indexed only if something new is reported for the substance in the original article | Particularly, in the introduction section, many original articles may summarize previously known information. However, unless something *new* is reported for the substance, the CAS Registry Number will not be indexed | At times, it may be advisable to search for some common names of substances as well as for the CAS Registry Numbers |
| The substance is indexed as precisely as possible | For example, the CAS Registry Number for morphine will be listed if the original article refers to morphine, but the CAS Registry Number for morphine sulfate will be listed if the article refers to the sulfate. Similarly, the CAS Registry Number for potassium will be listed if the article refers to potassium (e.g. potassium levels in blood), but for potassium ion if the article refers to $K^+$ | In some cases, a few CAS Registry Numbers (e.g. the parent base and all its salts) may be needed to cover a 'substance' |
| Since 1985, if a simple name for the substance is given by the author, it may be included after the CAS Registry Number, although no attempts are made to apply systematic nomenclature | For example, 'oxygen' is added after the CAS Registry Number 7782-44-7 in Figure 1.3, but the other CAS Registry Numbers refer to substances with complex names and no name term is added | CAS Registry Numbers should nearly always be used in search terms for substances |
| If the original report refers to a number of substances of a certain class, then the substance class heading also is indexed | For example, the substance class heading 'Porphyrins' is used in Figure 1.3 | General information on substance classes may be searched effectively through the substance class index headings |



CAS Registry Numbers are precise and comprehensive search terms for substances and should nearly always be included in the search for substances. Search results should always be checked to see that the CAS Registry Number has been used.

The four CAS Registry Numbers in Figure 1.3 are 158012-18-1, 159858-86-3, 159858-87-4, and 7782-44-7, but only the last one has a nomenclature term. It is thus apparent that only searches using the first three registry numbers will retrieve the substances indexed in this record, which clearly has important information relating to these substances!

### *1.5.4 Index Terms: CAS Roles*

All CAS Registry Numbers and substance class headings are followed by CAS Roles (for the list of roles, see Appendix 2). Roles have been assigned by the indexers since October 1994, while roles in the records in CAPLUS before that time have been assigned 'algorithmically'. That is, the roles have been assigned by computer-based searches that involved combinations of searches in section codes, controlled terms, and keywords. Note that searching with CAS Roles is a precision tool and is very useful for focusing on specific studies, particularly when large numbers of records occur for individual substances. The use of CAS Roles in searching through SciFinder Scholar is described in Chapter 5, Section 5.2.

### *1.5.5 Index Entries: Text-Modifying Phrases*

Text-modifying phrases are terms in parentheses that follow the index headings or CAS Registry Numbers. They are often terms from the original article that relate most directly to the index heading, so may be considered as author-related terms that qualify the index heading. The inclusion of these terms is significant since one of the features of SciFinder Scholar is that, *inter alia*, the user may choose answers in which the concepts searched are 'closely associated'. In general, SciFinder Scholar defines terms to be 'closely associated' when they appear in the title, in a single sentence in the abstract, or within a single index term and its text-modifying phrase. The assumption is that terms that are 'closely associated' (rather than anywhere in the reference) are more directly related, and so the inclusion of text-modifying phrases within the index heading provides an important level of precision in the choice of answer sets.

In cases in which text-modifying phrases are common to a number of index headings, the index headings are grouped and the single text-modifying phrase is applied. So, in Figure 1.3, the text-modifying phrase '(sapphyrins

spectroscopy and photosensitization in solns. and biol. membranes)' relates to (i.e. is 'closely associated' with) *each* of the six index headings before it.

### *1.5.6 Index Entries: Subheadings*

The first part of the text-modifying phrase may contain a subheading. For example, in Figure 1.3, the entries 'Porphyrins, sapphyrins' and 'Oxygen, singlet' appear. At present, it is sufficient simply to note that these are subheadings, although as further functionality is introduced into SciFinder Scholar, these subheadings may become significant as they form part of the index hierarchy.

Sometimes, there may be a number of subheadings and users initially may not completely understand the entry. Since subheadings are entered first in the text-modifying phrase, it may be necessary to read the sentence in a different way. So, the entry:

> IT 7782-44-7, Oxygen, biological studies
> RL: BIOL (Biological study)
> (singlet, formation of, porphyrins photosensitization of, after laser irradn. in aq. and alc. solns.)

is to be 'read' as 'formation of singlet oxygen by porphyrin photosensitization after laser irradiation in aqueous and alcoholic solutions'.

### *1.5.7 Index Entries: Supplementary Terms*

Supplementary terms are natural language words or phrases that are entered by the indexers to provide key additional information about the content of the document. Generally, they relate to the authors' terminology, and although they are not controlled vocabulary terms, they may be particularly useful for refining answer sets (see Chapter 2, Section 2.5.4).

In a manner somewhat similar to that in which text-modifying phrases are often author-related terms that expand on the index term, supplementary terms may be considered as index-related terms that expand on the terms used by the authors in titles. So, supplementary terms and terms in titles are relatively precise terms.

A supplementary term in Figure 1.3 is 'biol' and, indeed, it is noted that throughout Figure 1.3, there are a number of abbreviated terms. It is very important to search for abbreviations, but as discussed in Chapter 2, Section 2.3.2, SciFinder Scholar automatically searches for the abbreviations when the corresponding full term is entered in the search!



Indexing is subject to a number of policies, and users may learn basic policies simply by looking at and interpreting records. What is important to realize is that the records are written in part by authors and in part by indexers, and the search strategies should take advantage of both the parts. The implications of this are discussed in Chapter 2.

## 1.6 NLM Bibliographic Database (MEDLINE®)

The MEDLINE database contains more than 12 million references to journal articles in life sciences with a concentration on biomedicine. It covers 4300 journals published since 1958 (*http://www.nlm.nih.gov/pubs/factsheets/jsel.html*) and is updated four times a week. A Fact Sheet, which summarizes the content of the database, is available at *http://www.nlm.nih.gov/pubs/factsheets/medline.html*.

An example of the content of a single record is shown in Figure 1.5. After the title and the bibliographic information, most records contain the abstract as presented in the original article. Next follows the indexing. A key to indexing in MEDLINE is the Medical Subject Headings (MeSH®), which is outlined in another Fact Sheet – *http://www.nlm.nih.gov/pubs/factsheets/mesh.html*. There are more than 19 000 main headings that are constructed into a thesaurus that links broader, narrower, and related terms in the hierarchy.

The structure of the thesaurus is detailed elsewhere – *http://www.nlm.nih.gov/mesh* – and the full thesaurus may be downloaded from *http://www.nlm.nih.gov/mesh/filelist.html*. However, a quick appreciation of what is involved may be seen through Figure 1.6, which shows the thesaurus for the main heading 'Staphylococcus' in MEDLINE on the STN network. In particular, note the BT at different hierarchical levels (Broader Term 1 Micrococcaceae, Broader Term 2 gram-positive Cocci etc.), the NT (Staphylococcus aureus and Staphylococcus epidermidis), and the definition of the main heading.

SciFinder Scholar currently does not allow searches based on MEDLINE thesaurus capabilities, but scientists may download the thesaurus and use it as a guide for searching broader or narrower terms. Certainly, MeSH® terms provide important entries into related documents, and experienced users of SciFinder Scholar will get enhanced recall if they are familiar with the terms related to their research area.

In a manner similar to the way CAS Roles qualify the CAS Registry Numbers and substance index headings in CAPLUS, most index terms in

**Bibliographic Information**

**Spectroscopy and photosensitization of sapphyrins in solutions and biological membranes.** Roitman L; Ehrenberg B; Nitzan Y; Kral V; Sessler J L Department of Physics, Bar Ilan University, Ramat-Gan, Israel. PHOTOCHEMISTRY AND PHOTOBIOLOGY (1994 Nov), 60(5), 421-6. Journal code: P69. ISSN:0031-8655. United States Journal; Article; (JOURNAL ARTICLE) written in English. AN 95098899 MEDLINE

**Abstract**

A spectroscopic and photophysical study of three new sapphyrin molecules is presented. The sapphyrin backbone that was derivatized to make them water soluble possesses an absorption band around 700 nm, a desired property for biological photosensitization. We studied the absorption and fluorescence spectra, from which evidence for aggregation in solvents of different polarities was obtained. The extent of aggregation is correlated with the nature of the attached moiety. The absolute quantum yields of singlet oxygen production were measured, with 1,3-diphenyl isobenzofuran as a model target, and were 0.13-0.18 in ethanol. The binding constants to liposomes and to cells were determined spectroscopically and were found to correspond to the hydrophobicities of the compounds, with an additional effect, ascribed to the sugar moiety, which was found in the case of one of the sapphyrins. The efficiency of photodamage to Staphylococcus aureus by sapphyrins and hematoporphyrin was equivalent, on the basis of cells killed per microgram of sensitizer in the incubation mixture.

**Controlled Terms**

Check Tags: Support, Non-U.S. Gov't; Support, U.S. Gov't, P.H.S.

Escherichia coli: ME, metabolism
*Liposomes
Oxygen
*Photosensitizing Agents: CH, chemistry
Photosensitizing Agents: ME, metabolism
*Porphyrins: CH, chemistry
Porphyrins: ME, metabolism
Solutions
Spectrometry, Fluorescence
Spectrophotometry, Ultraviolet
Staphylococcus aureus: ME, metabolism

**Registry Numbers**

17778-80-2 (singlet oxygen)
7782-44-7 (Oxygen)

**Chemical Names**

0 (Liposomes)
0 (Photosensitizing Agents)
0 (Porphyrins)
0 (Solutions)

**Figure 1.5** Sample record from MEDLINE.

MEDLINE are qualified, and the qualifiers present in records for the heading 'Staphylococcus' are given in the AQ field in Figure 1.6. A full list of allowable qualifiers is available at *http://www.nlm.nih.gov/mesh/topcat.html*, although to assist users the qualifiers and the acronym are always present in the actual database record. For example, in Figure 1.5, the heading 'Photosensitizing Agents' is qualified both by CH (chemistry) and ME (metabolism).

In addition to the index headings and allowable qualifiers, the indexers add an asterisk (*) to those index headings considered to be key terms related to the article. Records in MEDLINE also may contain CAS Registry



```
=> E STAPHYLOCOCCUS/CT+ALL
E1            0    BT5    B Organisms/CT
E2        50112     BT4    Bacteria/CT
E3         3568      BT3    Gram-Positive Bacteria/CT
E4          430       BT2    Gram-Positive Cocci/CT
E5          201        BT1    Micrococcaceae/CT
E6        16866         -->    Staphylococcus/CT
E7        37865         MN     B3.510.400.500.846./CT
                         DC     an INDEX MEDICUS major descriptor

NOTE A genus of gram-positive,facultatively anaerobic,coccoid
bacteria. Its organisms occur singly, in pairs, and in tetrads and
characteristically divide in more than one plane to form irregular
clusters. Natural populations of Staphylococcus are membranes of
warm-blooded animals. Some species are opportunistic pathogens of
humans and animals.
INDX  infection = STAPH INFECTIONS on data form;
      staphylococcal clumping factor = COAGULASE (see X refs there);
      DF: STAPH
AQ     CH CL CY DE EN GD GE IM IP ME PH PY RE UL VI
MHTH   NLM 1966

E8            0          UF     STAPH/CT
E9        19955          NT1    Staphylococcus aureus/CT
E10        2714          NT1    Staphylococcus epidermidis/CT
```

**Figure 1.6** Example of the thesaurus in the MEDLINE file on STN. The thesaurus gives broader terms (BT), narrower terms (NT), a definition for the term (NOTE), the allowed qualifiers (AQ), and used for terms (UF), and thus suggests many search opportunities.

Numbers, chemical names, and chemical terms. Note that the CAS Registry Numbers in Figure 1.5 do not correspond exactly with those used for the same original article in CAPLUS (Figure 1.3), but this simply reflects the different indexing policies of the two organizations. Although CAPLUS contains more than 33 million CAS Registry Numbers, only slightly more than 56 000 of these have listings in MEDLINE. The implications of this are discussed in Chapter 5, Section 5.2.

This section is intended to give searchers just an outline of MEDLINE indexing and to point to additional resources that may help those who wish to use SciFinder Scholar at a very advanced level. In summary, it is generally sufficient for users to understand that an index entry:

*Photosensitizing Agents: CH, chemistry

indicates that the MeSH® heading is 'Photosensitizing Agents', the * indicates a key heading for the record, and 'CH, chemistry' is a qualifier that more specifically links the nature of the research to the MeSH® heading.

## 1.7 CAS Substance Database (REGISTRY)

While bibliographic databases have text from authors and indexers, chemical substance and chemical reaction databases are produced entirely by indexers. Consequently, it helps if searchers understand the index policies that have been applied, and some of these are discussed below (and in more detail in Chapters 3 and 6).

Up-to-date information on REGISTRY is available from *http://www.cas.org/casdb.html*. The database currently contains more than 47 million substance records, which include over 21 million organic and inorganic substances and over 26 million sequences. An example of a record in SciFinder Scholar is shown in Figure 1.7.

After the CAS Registry Number, the current CAS systematic name is given, followed by former CAS names, common usage names, and trade names. These complete names may be searched through SciFinder Scholar (see Chapter 3, Section 3.6 and Chapter 5, Section 5.2). The print version of Chemical Abstracts is indexed in five-year collective indexes and the terms 9CI and 8CI in the name fields in Figure 1.7 refer to the ninth and eighth Collective Index, respectively. They are of no significance with respect to search strategies.

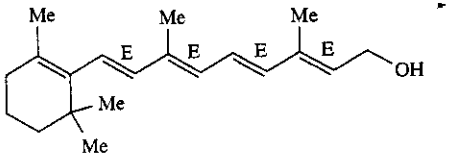
Most substance records in REGISTRY list files on the STN network that contain the CAS Registry Number for the substance, and at times, these specific databases may yield additional information. The files automatically included in SciFinder Scholar are CAPLUS (the CAS bibliographic database), CASREACT (the CAS chemical reaction database), CHEMCATS (the CAS commercial substances database), CHEMLIST (the CAS regulated chemicals database), and MEDLINE (the NLM bibliographic database). When the substance is commercially available or is included in the regulated chemicals list, the SciFinder Scholar record for the substance contains a link to the record in the appropriate database.

The indexing of substances is subject to a number of issues, which may apply either to the class of substance (e.g. salts, polymers, alloys, mixtures) or to the nature of the bonding in the molecule (e.g. issues such as resonance, $\sigma$- and $\pi$-bonding need to be addressed). SciFinder Scholar automatically handles most of these issues, although there are many complications and the user is well advised to study the basic aspects of the indexing of substances. With the aid of examples, Appendix 5 explains the most commonly encountered issues.



**Registry Number:** 68-26-8
**CA Index Name:** Retinol (9CI)
**Other Names:** Retinol, all-trans- (8CI); (all-E)-3,7-Dimethyl-9-(2,6,6-trimethyl-1-cyclohexen-1-yl)-2,4,6,8nonatetraen-1-ol; $\beta$-Retinol; 2,4,6,8-Nonatetraen-1-ol, 3,7-dimethyl-9-(2,6,6-trimethyl-1-cyclohexen-1-yl)-, (all-E)-; A-Mulsal; A-Vi-Pel; Acon; Afaxin; Agiolan; Agoncal; Alcovit A; all-trans-Retinol; all-trans-Retinyl alcohol; all-trans-Vitamin A; all-trans-Vitamin A alcohol; all-trans-Vitamin A1; Alphalin; Alphasterol;. . ... (*additional names follow*)
**Formula:** C20 H30 O
**STN Files:** CAPLUS, AGRICOLA, ANABSTR, BEILSTEIN, BIOBUSINESS, BIOSIS, BIOTEHNO, CA, CABA, CAOLD, CASREACT, CBNB, CEN, CHEMCATS, CHEMINFORMRX, CHEMLIST, CIN, CSCHEM, CSNB, DDFU, DETHERM, DRUGU, EMBASE, GMELIN, HSDB, IFICDB, IFIPAT, IFIUDB, IPA, MRCK, MSDS-OHS, NAPRALERT, NIOSHTIC, PHAR, PIRA, PROMT, RTECS, SPECINFO, TOXLINE, TOXLIT, ULIDAT, USAN, USPATFULL, VETU
(Additional Information is available through STN International. Contact your information specialist, a local CAS representative, or the CAS Help Desk for Assistance)
**Deleted Registry Number(s):** 13123-33-6, 17104-91-5, 5979-23-7

Double bond geometry as shown.




Commercial Sources
Regulated Chemicals Listing
**~5732 References**

**Figure 1.7** Sample record from REGISTRY. (Note: only some of the alternative names are given in this display.) Copyright the American Chemical Society and reproduced with permission.

## 1.8 CAS Regulatory Information Database (CHEMLIST®)

In order for national authorities to have a mechanism to regulate trade in chemicals, many countries require companies to register the substances

**Accession Number:** 323 CHEMLIST
**CAS Registry Number:** 68-26-8
**Chemical Name**
Retinol (English, French, German, Spanish) (TSCA, DSL, EINECS, AICS)
Vitamin A (ENCS)
Vitamin A alcohol (ENCS)
(all-E)-3,7-Dimethyl-9-(2,6,6-trimethyl-1-cyclohexene-1-yl)-2,4,6,8-nonatetraen-1-ol (ECL)
(all-E)-3,7-Dimethyl-9-(2,6,6-trimethyl-1-cyclohexen-1-yl)-2,4,6,8-nonatetraen-1-ol
β-Retinol
2,4,6,8-Nonatetraen-1-ol, 3,7-dimethyl-9-(2,6,6-trimethyl-1-cyclohexen-1-yl)-, (all-E)-
A-Mulsal ....

**File Segment**
AUSTRALIA: AICS; CANADA: DSL; EEC: EINECS; JAPAN: ENCS; KOREA: ECL; USA: FDA, STATE, TSCA

**Confidentiality Status**
Public
**Regulatory List Number**
EINECS No.: 200-683-7
ENCS No.: 9-1029; 9-1053
ECL Serial No.: KE-11884

**Inventory Status**

| | |
|---|---|
| On TSCA Inventory | January 2000 Inventory Tape. |
| On DSL | Supplement to Canada Gazette, Part I, January 26, 1991. |
| On EINECS | Annex to Official Journal of the European Communities, 15 June 1990. |
| On ENCS | Japanese Gazette. Contained within class: Low Molecular Carbo-monocyclic Organic Compounds. |
| On AICS | Australian Inventory of Chemical Substances, June 1996 Ed. |

**FDA Priority-Based Assessment of Food Additives**
FDA Priority-Based Assessment of Food Additives
Priority-Based Assessment of Food Additives (PAFA) File, FDA Center for Food Safety and Applied Nutrition (CFSAN) (1998)
Listed Name(s): Vitamin A

**==== Miscellaneous Regulations ====**

**Hazard, Toxicology, and Use Information**
Hazard, Toxicology, and Use Information
Human Data
Mutation data
Reproductive Effect (RTECS)

**==== U. S. State Regulations ====**
**Massachusetts Right-to-Know**
Massachusetts Right-to-Know
Massachusetts Substance List for Right-to Know Law (11 Apr 94); General Law C.111F, Chapter 30A (28 Jun 84); 105 CMR 670.000, Appendix A
Listed Name(s): Vitamin A

**Figure 1.8** Sample record from CHEMLIST. Copyright the American Chemical Society and reproduced with permission.



before their manufacture or distribution. Such inventories are essential, for example, for the monitoring of illegal substances and for keeping track of environmental issues relating to chemicals. The primary reference point in these national inventories is the CAS Registry Number.

The CAS has built a database of regulated chemical substances from a number of national and international chemical inventories and regulatory lists. The database contains more than 230 000 chemical substances and is updated weekly. Details of the content are available at *http://www.cas.org/CASFILES/chemlist.html* and a typical record is shown in Figure 1.8.

It is neither possible nor necessary to search directly in CHEMLIST with SciFinder Scholar. Instead, the user first needs to find the chemical substance and then click on the link labeled 'Regulated Chemicals Listing'. In general, the staff and students at universities are not involved in registering chemicals with national authorities, although safe laboratory practices apply to university laboratories and the regulated substance database gives important references to legal requirements and safety issues.

## 1.9 CAS Chemical Catalog Database (CHEMCATS®)

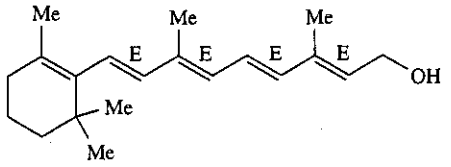
The CAS chemical catalog database *http://www.cas.org/CASFILES/chemcats.html* contains more than 5 million records from approximately 660 suppliers and 790 catalogs. Each record contains the catalog information for the substance (e.g. chemical and trade names, the company names and addresses) and the supplier information (e.g. pricing terms) and an example is shown in Figure 1.9.

Again, it is neither possible nor necessary to search directly in CHEMCATS. Instead, the strategy is to find the substance first. When a substance appears in one of the chemical catalogs, a link is provided (e.g. the link 'Commercial Sources' in Figure 1.7) to the chemical catalog database and it is easy for the user to scroll through the company information to identify the local supplier.

## 1.10 CAS Chemical Reaction Database (CASREACT®)

The reaction database contains over 6 million reactions selected from journal articles since 1907 – *http://www.cas.org/CASFILES/casreact.html*. A typical example in SciFinder Scholar is shown in Figure 1.10. In the reaction database, all the atoms and bonds are correlated between starting material and product, and bonds being formed or broken are tagged; however,

**Catalog Name:** SIGMA
**Publication Date:** 25 Feb 2000
**Order Number:** R7632
**Chemical Name:** all trans-Retinol
**Synonym:** Vitamin A
**Registry Number:** 68-26-8




**Pricing:**
Quantity: 1 25 MG, Price: inquire
Quantity: 1 100 MG, Price: inquire
Quantity: 1 250 MG, Price: inquire
Quantity: 1 500 MG, Price: inquire
Quantity: 1 1 G, Price: inquire
Notes: 1999

**Company Info:**
SIGMA
P O Box 14508
St. Louis, MO, 63178-9916 USA
Phone: 1 800 325 3010
314 771 5750
Fax: 1 800 325 5052
314 771 5757
Email: sigma@sial.com
Web: http://www.sigma-aldrich.com

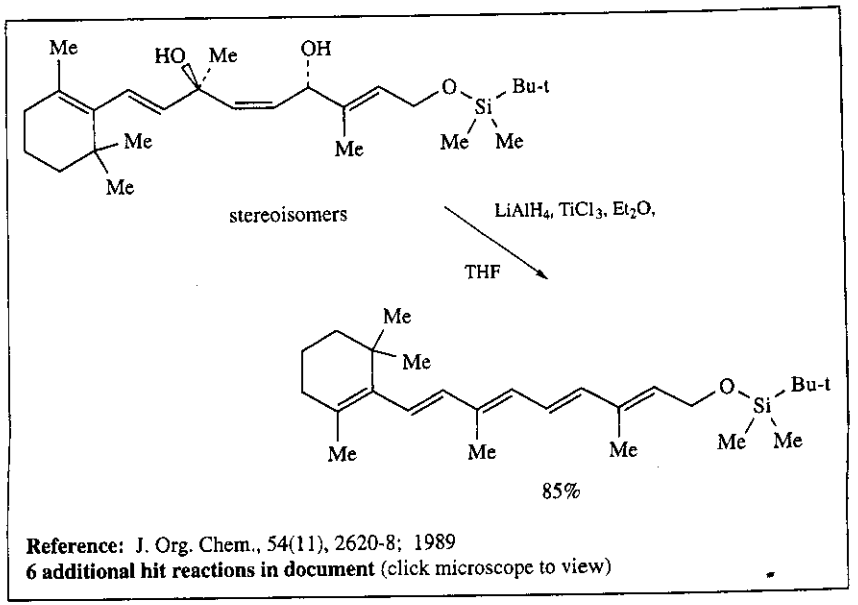
**Company Info:**
SIGMA-ALDRICH de Argentina, S.A.
Av. Pueyrredon 2446/50,Piso 5-B
Buenos Aires, 1119 Argentina
Phone: 54 1 807 0321
Fax: 54 1 807 0346
Email: rcsjr@compuserv.com

*(Included also is information from other offices worldwide)*

**Figure 1.9** Sample record from CHEMCATS. Copyright the American Chemical Society and reproduced with permission.

these details are not seen in the reaction diagram as shown (Figure 1.10). This enables precise reactions to be retrieved easily, and this is important since the scientist may not only wish to know that a substance has been prepared but may also wish to know the preparations involving the formation of specific bonds.






**Reference:** J. Org. Chem., 54(11), 2620-8; 1989
**6 additional hit reactions in document** (click microscope to view)

**Figure 1.10** Sample record from CASREACT. Copyright the American Chemical Society and reproduced with permission.

It is helpful to appreciate that only key or representative new reactions from the original documents are fully indexed and that the database includes data from 1907. Even so, it is one of the world's premier reaction databases and generally produces many more relevant answers than alternative desktop reaction databases. SciFinder Scholar offers a number of other options for searching for chemical reactions, and further details are discussed in Chapter 6.

## 1.11 Exercises

1.1 Spot the difference! In Figures 1.3 and 1.5, what differences are there in the:

(a) bibliographic parts of the records;

(b) abstract;

(c) subject indexing; and

(d) substance indexing.

1.2 Why is the substance 1,3-diphenylisobenzofuran not mentioned in the indexing in Figures 1.3 and 1.5, even though it is in the abstract? What are the implications?

1.3 Which part of the record in Figure 1.3 is written by the indexer?

1.4 What policies do authors apply when writing titles and abstracts, and what should authors be keeping in mind when they write these sections?

1.5 Discuss the advantages and disadvantages of searching those parts of the records written by:

(a) the author;

(b) the indexer.

1.6 Figure 1.3 contains four CAS Registry Numbers, but only one has a name associated. Why do CAS use Registry Numbers as index entries for substances and what are the implications of this for the searcher?

1.7 On the basis of the information in Figures 1.3 and 1.5, what search terms would be needed to find all records that:

(i) contained the concept photosensitization;

(ii) are written by Jonathan L. Sessler;

(iii) are listed for Bar Ilan University; and

(iv) involve the substance singlet oxygen.

(HINT: To answer this question, it is sufficient to write down the actual terms that are in the records.)

1.8 One of the search options in SciFinder Scholar is to restrict entries to those where the search terms are 'closely associated' (which usually means just within the title, within one sentence in the abstract, or within the one index term including its text-modifying phrase). While each of the following words appear somewhere in the record in Figure 1.3, which combinations of any two terms appear only 'closely associated'? What are the implications for the searcher?

| | | |
|---|---|---|
| sapphyrins | porphyrins | singlet oxygen |
| photosensitization | laser radiation | Staphylococcus aureus |
| hematoporphyrin | 7782-44-7 | |

1.9 The CAS Registry Number for hematoporphyrin is 14459-29-1. The word hematoporphyrin is in the abstract in Figures 1.3 and 1.5, but the



CAS Registry Number is not indexed. Why is this so, and what are the implications for searches on the substance?

1.10 In Figure 1.7, the substance 'retinol' lists a number of STN files in which the CAS Registry Number appears but MEDLINE does not appear. Since retinol is an important substance involved with the chemistry of vision, is it not surprising that the CAS Registry Number 68-26-8 is not in MEDLINE? What is the 'problem' here? (HINT: unless you already have good knowledge of the science and of the databases, you will not be able to answer this question at this stage! However, when you logon to SciFinder Scholar, you may wish to find the CAS Registry Numbers 68-26-8 and 11103-57-4 and then think about what is happening!)

Chapter 2

# Explore by Research Topic

## 2.1 Introduction

Information on topics is obtained through the CAPLUS and the MEDLINE databases. First the searcher clicks **Explore by Research Topic** (Figure 2.1), then enters the query (Figure 2.2), and a list of topic candidates is obtained (Figure 2.3).

Next, the user may go back to enter another phrase or may proceed to choose the appropriate candidates, in which case, the reference screen (Figure 2.4) is obtained. From this point, the user has many options, but a next step may be to click the microscope adjacent to the record of interest. Information of the type given in Figures 1.3 and 1.5 is displayed, and depending on the availability in the user's institution, the full text document may immediately be recalled.

None of this requires any knowledge of how the search was performed or of the databases, and very often, good answers are obtained immediately with minimum effort. SciFinder Scholar is extremely easy to use at the basic level, and the scientist should always remember that some of the best results in science come from the simplest experiments. So, a searcher who is unsure of how to approach a problem should just try something!

However, how does SciFinder Scholar convert the initial query to concepts (i.e. what is 'behind the scenes' in going from Figure 2.2 to Figure 2.3?) and what has been searched to give the number of candidates? What are the issues in the choice of the appropriate candidates from the list (Figure 2.3)

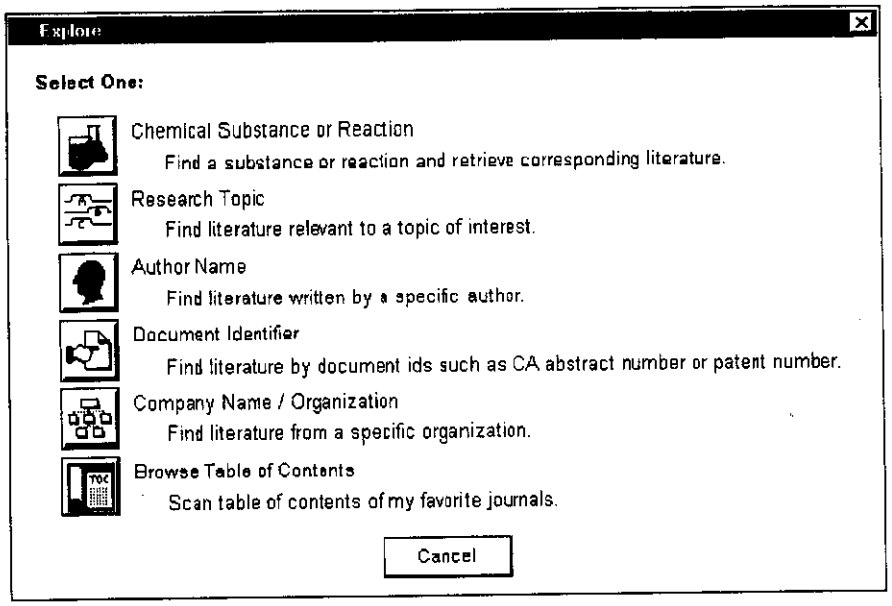



**Figure 2.1** The initial explore screen in SciFinder Scholar offers many ways to commence the search. (The initial screen for SciFinder is shown in Figure 4.13.) SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

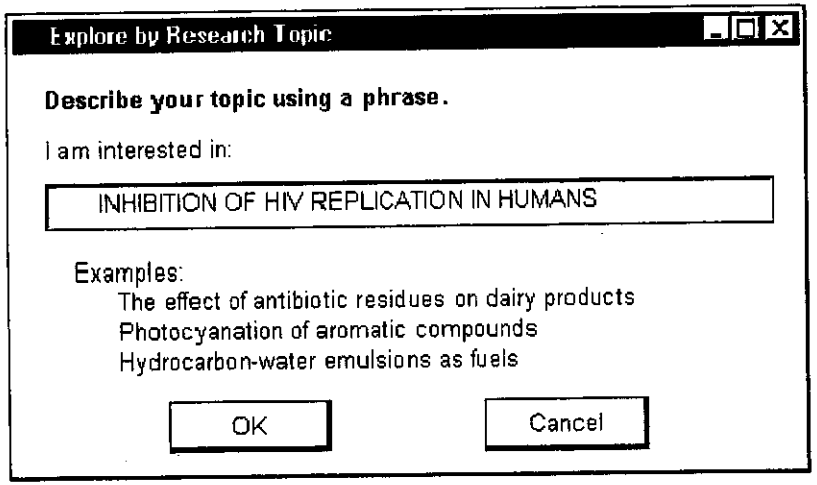



**Figure 2.2** **Explore by Research Topic** screen. Terms are entered, preferably with prepositions between the key concepts. SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.



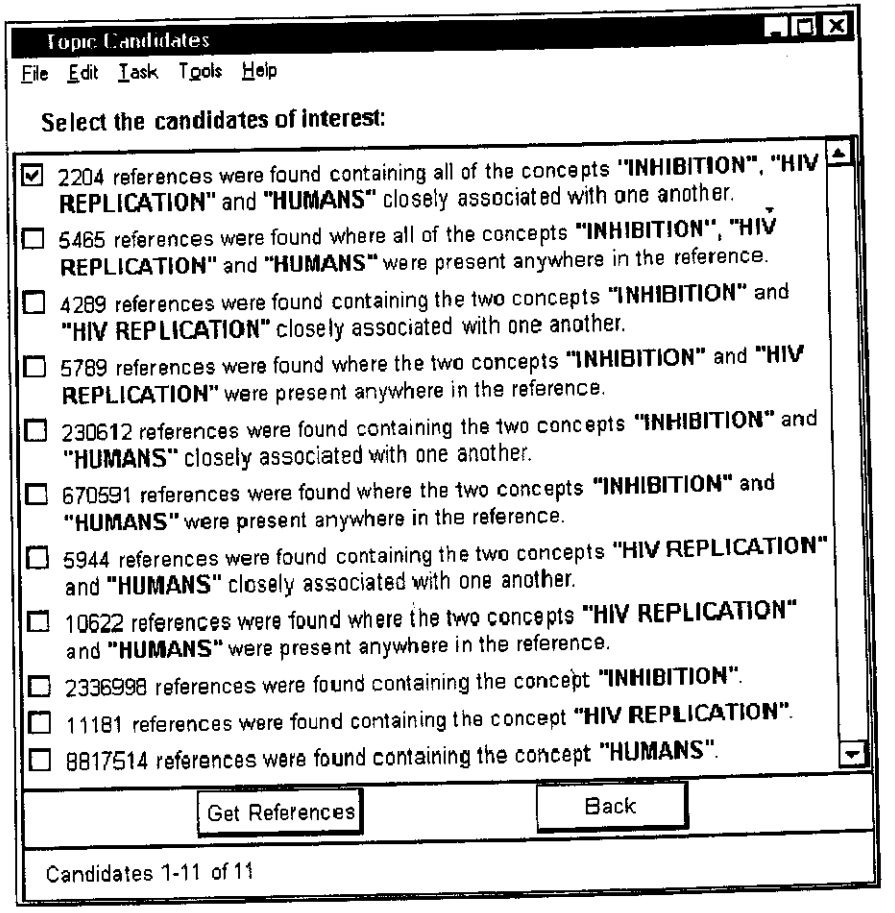



**Figure 2.3** List of candidates following entry of terms under **Explore by Research Topic**. The concepts are identified in bold, and the number of records from the combinations of the concepts are indicated. SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

and which features from the reference screen (Figure 2.4) should be used and when? Perhaps, most importantly, how may an understanding of these issues help the scientist increase comprehensiveness and precision in answers and assist in the production of creative solutions to complicated problems?

This chapter addresses these questions!

By default, **Explore by Research Topic** searches for terms in the title, abstract, and index fields in both databases. Searches may be restricted to CAPLUS by choosing only this option under the pull-down header

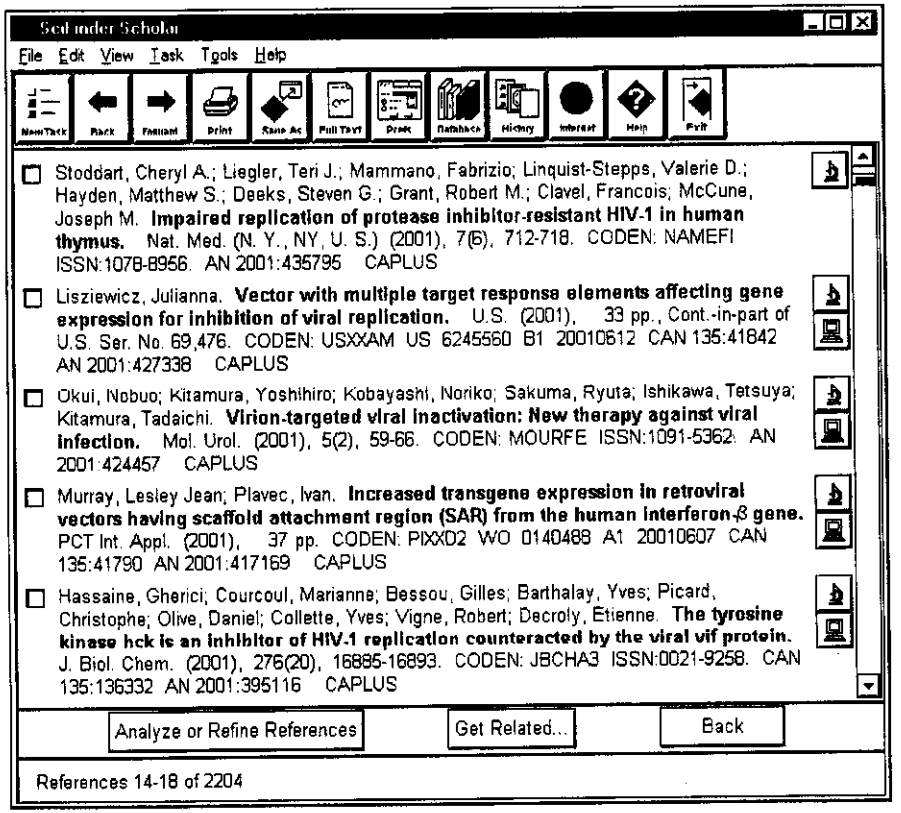



**Figure 2.4** The initial reference screen gives bibliographic information and title. In the default display, records from CAPLUS appear before records for MEDLINE, although alternative orders may be chosen through **View**. Copyright the American Chemical Society and reproduced with permission.

**Databases** before the search is conducted. Actually, there is little need to do this since searching both databases takes advantage of the different indexing and the different database coverage and thus increases recall.[1] Once the answers have been obtained, the number of records in the different databases may be determined (Section 2.5.4). It is far better to first see what each database provides and then make a decision on how to proceed, rather than to preempt a result in one database and potentially miss out on important answers.

[1] Task package users of SciFinder should note that searching MEDLINE in addition to CAPLUS involves an additional 0.5 task.



## 2.2 How SciFinder Scholar Converts the Query to a List of Candidates

SciFinder Scholar uses an algorithm that first identifies the individual concepts (Section 2.3) through the prepositions and conjunctions entered in the query. While the actual preposition (of, with, at, etc.) is of no significance, the conjunction chosen (and, not, or) may be critical. Certain stop/controlled words such as 'the effect', 'information on', and so forth are also used to identify concepts and may not be searched. Stop/controlled words may easily be recognized as they are not included within the concepts shown in the list of candidates.

### *2.2.1 Search Fields*

Next, SciFinder Scholar searches for the concepts in the title, abstract, and index fields and may produce answer candidates where all the concepts are 'closely associated' and 'anywhere in the reference'. The entry 'closely associated' usually means that all the concepts are in the title, in the same sentence in the abstract or in a single index heading field (Chapter 1, Section 1.5.5), while 'anywhere in the reference' means that the concepts are present somewhere in the title, abstract, and index fields. The usual assumption is that the closer the concepts are, the more directly related they are, and so the former answer set may afford a greater level of precision.

### *2.2.2 Candidates*

SciFinder Scholar may then display candidates where combinations of *some* of the concepts are 'closely associated' or 'anywhere in the reference', and finally, displays the candidates that represent answers for the individual concepts (Figure 2.3). The number of candidates displayed ultimately depends on the number of concepts identified and whether or not conjunctions (Boolean operators) have been included in the question. Some examples are displayed in Table 2.1.

*The first task is to look very carefully at the concepts that have been identified in order to verify that the algorithm has been applied in the intended manner.* In the example chosen (Figure 2.3, and Entry 3 in Table 2.1), the concepts identified are 'inhibition', 'HIV replication' and 'human', and the result is a list of 11 candidates where the concepts are variously listed 'closely associated' or 'anywhere in the reference'.

However, if the question is asked in a slightly different way, different concepts may be identified and different numbers of candidates may be

**Table 2.1** Examples of the number of candidates identified from different Explore entries

| Entry | Query: I am interested in | Number of concepts | The concepts identified | Number of candidates usually listed |
|---|---|---|---|---|
| 1 | Inhibition of replication of HIV in humans | 4 | 'Inhibition', 'HIV', 'replication', 'humans' | 26 |
| 2 | Inhibition and replication and HIV and humans | 4 | 'Inhibition', 'HIV', 'replication', 'humans' | 6 |
| 3 | Inhibition of HIV replication in humans | 3 | 'Inhibition', 'HIV replication', 'humans' | 11 |
| 4 | Inhibition and HIV replication and humans | 3 | 'Inhibition', 'HIV replication', 'humans' | 5 |
| 5 | Inhibition of human HIV replication | 2 | 'Inhibition', 'human HIV replication' | 4 |
| 6 | Inhibition HIV replication humans | 2 | 'Inhibition HIV', 'replication humans' | 3 |

listed (Table 2.1). For example, because of the extra preposition in Entry 1 (Table 2.1), 4 concepts are identified and they and their various combinations are listed as 'closely associated' or 'anywhere in the reference' to give 26 candidates. On the other hand, Entry 2 (Table 2.1) in which the terms are connected with the 'and' conjunction has only six candidates. The essential difference is that the use of 'and' here is strictly interpreted so that the terms are identified only as 'anywhere in the reference'; the 'closely associated' candidates are not displayed. Other entries in Table 2.1 are similarly explained, although if four terms are entered consecutively (Entry 6, Table 2.1), SciFinder Scholar automatically may break them into two concepts. The reason is that it is not advisable to search for several words in a single concept, and SciFinder Scholar tries to guide the user here by breaking up several words (which are grouped together) into different 'concepts'.

Conjunctions 'and, not, or' may be entered in the query but once again the list of candidates needs to be studied carefully. Part of the issue here is that it is better for the searchers to use 'and' and 'or' as precisely as possible. However, in recognition that searchers may at times use 'and' when 'or' is intended, SciFinder Scholar sometimes interprets the conjunction 'and' in the more general sense. It all depends on the precise query. For example, although many of the candidates listed after 'I am interested in ''treatment of HIV in men and women'' ' are similar to the candidates listed after 'I am interested in ''treatment of HIV in men or women'' ' (i.e. 'and' is interpreted in part as 'or' – however, it is better to use 'or' rather than 'and' for synonyms), there are some unique candidates in each set.



At all times, care must be exercised in the use of the conjunction 'not' since relevant records may be excluded. Indeed, in general, use of the word 'not' is not advised and it is preferable to use some of the alternative strategies mentioned in this chapter to make answer sets more precise.

### *2.2.3 Notes on Terms Entered*

There are a number of additional points to note. First, at times users enter too many concepts, and more appropriate answer sets may be retrieved when fewer concepts are employed. While all the concepts may be in the original publication, the records being searched include only title, abstract, and index entries in which all the concepts may not be mentioned. Through the display of combinations of the concepts, SciFinder Scholar simply is guiding the searcher, who, for example, in cases in which very few records mention all the concepts, may instead be alerted and may choose candidates with fewer concepts before exploring the question further.

Second, the inclusion of several prepositions (Entry 1, Table 2.1) produces the greatest number of candidates, and these candidates will ultimately include the options displayed in Entries 2 to 6 (Table 2.1). Having more options is helpful since the user has a better idea beforehand of the types of answer sets that may be of interest. However, if more than four concepts are identified, SciFinder Scholar may faithfully carry out instructions, and the lists may be tedious to work through (e.g. the permutations and combinations of five concepts linked with prepositions gives 57 candidates).

Third, the number of records for the individual concepts at the bottom of the list provides an important piece of information. For example, a search 'I am interested in ''removal of cyanide from wastewater from gold mine tailings'' ' produces only one and seven records in which all the concepts are 'closely associated' and 'anywhere in the reference', respectively. This is clearly a very small number of hits for such a topic in a database that is particularly strong in the area of mining and ore extraction. The problem is evident when the number of hits for the individual concepts is examined, and in this case, there are only just over 200 hits for 'gold mine tailings'. So, this concept needs redefining.

The issue is that when two or three words are within the same concept, SciFinder Scholar will look for records containing *all* of them in the same sentence, but there may be sentences that have just one of the words by itself

or the word with quite different second (or third) words. So, while there are just over 200 hits for the concept 'gold mine tailings', there are more than 3000 hits for the concept 'gold mine' and more than 180 000 for the concept 'gold'. This last answer set includes words such as 'gold mine wastes' and 'gold processing plants' that are clearly relevant to questions that concern the removal of cyanide from wastewater in the gold industry. Accordingly, a search 'I am interested in ''removal of cyanide from wastewater with gold'' ' is better, and now more than 80 references contain all the four concepts 'closely associated'.

In summary, the preferred option is to enter between two to five terms in the initial query, to include prepositions between the concepts, and to avoid too many words within a single 'phrase'. Then, SciFinder Scholar suggests more alternatives to the searcher, and indeed, the ability of SciFinder Scholar to guide the searcher through prior alerting of alternatives is one of its greatest benefits to facilitate the searching of topics.

## 2.3 The Identification and Meaning of Concepts

It is essential to check the list of candidates to verify that SciFinder Scholar has identified the concepts as intended. With the possible exception of Entry 6, all the candidates in Table 2.1 are logical outcomes from the query. If the working of the algorithm in Entry 6 causes difficulties, then it is a simple matter to interpose a few prepositions or else restrict the consecutive terms to less than four words. For example, 'I am interested in ''fourier transform infrared spectroscopy'' ' will produce two concepts ('fourier transform' and 'infrared spectroscopy'), whereas 'I am interested in ''fourier transform infrared'' ' produces a single concept that effectively has the same meaning.

However, at times, the concepts identified will not be as intended, in which case, alternative entries need to be made under **Explore by Research Topic**. The most common problem lies with the issue of 'distributed modifiers', that is, with different terms that qualify another term. For example, it is perfectly acceptable to express in the English language, 'I am interested in liver or kidney diseases', but since SciFinder Scholar uses the word 'or' to identify concepts, the concepts identified in this query are 'liver' and 'kidney diseases'. The appropriate entry may thus be 'I am interested in ''liver diseases or kidney diseases'' ' where SciFinder Scholar identifies the concepts 'liver diseases' and 'kidney diseases' as intended.

Depending on the original entry, there may be a few other reasons for the algorithm not to interpret the query as intended. It is a simple matter to note



the concepts that SciFinder Scholar identifies and to make logical revisions to the original query where necessary.

The list of candidates (Figure 2.3) mentions 'concepts' that are determined by SciFinder Scholar after the application of a number of rules.

### *2.3.1 Automatic Truncation*

Truncation of search terms allows for the retrieval of words that contain a common word fragment, and SciFinder Scholar applies truncation automatically. This saves the user from having to think about truncation, but exactly where to truncate is a tricky problem. Truncation too late may exclude relevant terms, whereas truncation too early may retrieve irrelevant terms that happen to contain the word fragment.

In order to determine how the automatic application of truncation has been applied, it is necessary to look through the full records (hit terms in SciFinder Scholar are highlighted in blue for easy identification). However, even from the reference screen (Figure 2.4), it may be noted that SciFinder Scholar has retrieved hits for 'inhibitor' as well as the term 'inhibition', which was entered, and quick examination of other records reveals that 'inhibiting', 'inhibit', 'inhibitors', and so forth are hit terms.

In the great majority of cases, automatic truncation works very effectively. However, if the answers include references in which the truncation does not appear optimal, it may be better to work through answers manually or to use analyze/refine options (Section 2.5.4) rather than make drastic decisions at this stage.

It is not possible in this text to give examples that cover all the instances that may be encountered, but as an example, a search for 'tropical plants' will produce hits for 'tropical' and 'tropics' as required. However, it will also produce hits that contain 'tropanes' (a particular class of chemical substances often found in plants). So, at this stage it is sufficient to remember that truncation is a complex issue. The algorithm in SciFinder Scholar initially favors comprehension over precision, since it is assumed that it is better to allow users to make decisions themselves on relevancy of answers. It is better to know what is present rather than not to know what may have been missed.

Solutions to issues relating to automatic truncation are discussed later.

### *2.3.2 Synonyms*

While the original query included only the term 'HIV', Figure 2.4 indicates that 'human immunodeficiency virus' was included as a search term. This occurs because within SciFinder Scholar is a very extensive synonym

dictionary, so that when a term is entered, SciFinder Scholar–predetermined synonyms are additionally searched automatically. This dictionary usually includes singulars and plurals, so generally it does not matter which form of the word is entered in the query. The dictionary also allows for American and English spelling, so the concept 'flavour' retrieves hits with 'flavour', 'flavours', 'flavor', and 'flavors'.

It was noted in Figure 1.3 that CAS uses many standard abbreviations, and the complete list of abbreviations is available at *http://www.cas.org/ONLINE/standards.html*. SciFinder Scholar automatically includes standard abbreviations where appropriate in the terms searched within the concept.

However, exactly what constitutes a 'synonym' may depend critically on the context, so not always will SciFinder Scholar necessarily search terms exactly as intended. The synonym dictionary, which is constantly being modified in order to achieve an optimal balance between search comprehension and precision, is not available, so the only way to identify terms that are automatically searched is to look through the records retrieved.

Automatic searching of predetermined synonyms greatly assists the user, who nevertheless still should consider adding additional synonyms. This is achieved at the Explore level either by using the conjunction 'or' or by entering terms in parentheses. For example, the queries 'I am interested in ''HIV in humans (men, women)'' ' and 'I am interested in ''HIV in humans or men or women'' ' give very similar lists of candidates, although the list from the latter query gives some additional 'closely associated' options. The differences between the candidates, however, are generally of little consequence because both contain the important candidates where 'HIV' and any of 'humans/men/women' are 'closely associated' and 'anywhere in the reference'.

The various synonyms entered in the question are effectively treated as separate concepts, so if many synonyms are added, the list of candidates may be extensive. Searching for information generally requires compromises, particularly relating to comprehension versus precision; it helps if the searcher is alert to the various issues and to apply the most appropriate search for the information being sought.

### *2.3.3 Phrases*

Words not separated by prepositions or conjunctions or other reserved words are identified as a single concept. Answers are retrieved where these words are 'closely associated', and not just as the exact phrase. This is important because searching for exact phrases fails to retrieve answers where the order of the words is different or where other words are in between, and since these



factors occur very frequently, searches for phrases will almost invariably miss important answers. Accordingly, it makes no difference if 'traditional Chinese medicines' or 'Chinese traditional medicines' are entered under **Explore by Research Topic** since records are retrieved that contain the three terms at least somewhere in the sentence (irrespective of the order).

While SciFinder Scholar thus correctly searches for words within the sentence and not as a phrase, there may be instances in which precise phrases are required. To assist the searcher in these cases, SciFinder Scholar may start the list of candidates with an indication of the number of records 'as entered'. For example, the entry 'I am interested in ''traditional Chinese medicines'' ' gives two candidates:

> 343 references were found containing '**traditional Chinese medicines**' as entered
> 6532 references were found containing the concept '**traditional Chinese medicines**'

Answers in the former set contain exactly the words 'traditional', 'Chinese', 'medicines' in that order. In the latter set, the answers will additionally include words within a sentence ('closely associated') and words where the SciFinder Scholar algorithm discussed within this chapter has been applied (e.g. 'medicine' and 'medicines').[2]

At times, records retrieved under candidates 'as entered' will not *exactly* match the entry. For example, an entry 'I am interested in ''oil in water'' ' will retrieve 'oil with water' and 'oil water'. The reason is that SciFinder Scholar may allow up to one intervening word between the main terms when a preposition is entered in the query. However, this is a detail that experienced users of SciFinder Scholar will observe and understand as an outcome of the algorithm that has been implemented by the developers of the product. These developers have in-depth knowledge of the databases and have included in the algorithm the functions that they consider will provide the best answer sets on the basis of the queries entered.

One of the main uses of candidates 'as entered' is to allow the inclusion of words that SciFinder Scholar does not include in concepts. For example, the list of candidates to the query 'I am interested in ''off flavours in wines'' ' gives candidates only with the concepts 'flavours' and 'wines'. While references with these two concepts may be obtained and then examined in turn or

[2] By comparison, the numbers of answers 'as entered' for the searches 'traditional Chinese medicine', 'Chinese traditional medicine', and 'Chinese traditional medicines', are 1901, 394, and 138, respectively.

narrowed by strategies mentioned later in this chapter, another option may be to enter the query 'I am interested in ''off flavours'' '. The candidate 'as entered' will include exactly 'off flavours' and the references may be chosen and later narrowed to terms relating to wines.

However, remember that the candidate 'as entered' does not employ the SciFinder Scholar algorithm and so, *inter alia*, American/English spellings, truncation, and singulars/plurals are *not* searched. Since the databases have almost exclusively American English spellings, the better query is 'I am interested in ''off flavors'' '. It always helps if users employ 'scientific method' throughout the process, that is, a critical analysis of the results of the initial 'experiment' be conducted (and in this case, the user would immediately think about the relatively few answers retrieved for off flavours 'as entered').

### *2.3.4 CAS Registry Numbers*

The importance of searching for CAS Registry Numbers for substances was illustrated in Chapter 1, Section 1.5.3, so it is necessary that one of the first actions of the algorithm behind **Explore by Research Topic** is to determine whether any of the terms entered corresponds to the exact name of a substance in REGISTRY. If this is the case, SciFinder Scholar automatically searches the CAS Registry Number as one of the terms within the concept.

However, the key issue is that a name in the substance database is recognized, and this is likely to occur only for substances with common or simple names. Unless the user has an excellent knowledge of CAS nomenclature, in more complicated cases, the name entered in the query may not match exactly.

It is helpful first, to remember that the CAS Registry Number should nearly always be used in searches for substances and second, to check in records retrieved that the CAS Registry Number is highlighted in blue and hence that it has been used as a search term. Further issues concerning retrieval of information on substances are discussed in Chapter 5, Section 5.2.

## 2.4 Choosing Candidates

Once the most appropriate list of candidates has been obtained, the next task is to choose options from the list. Generally, some of the first listed candidates will contain more precise answers and will have fewer references as they will include the greater number of concepts. However, if there are very few records with all the concepts, then clearly, either the question



needs to be revised or else other candidates with fewer concepts need to be chosen.

On the other hand, if there are a large number of references with all the concepts, then either more concepts must be added to the question or else **Get References** for appropriate candidates should be obtained and further refinements undertaken at a later stage. Since SciFinder Scholar has many options for subsequent refinements (Section 2.5.4), the *latter option generally is better* and indeed users should not be deterred by initial answer sets of several thousands of records.

Candidates where the concepts are 'closely associated' will generally produce more precise answer sets and are often chosen where many references are obtained. Nevertheless, it is important to realize that 'closely associated' restricts answers to cases in which the terms are in the title, in the same sentence of the abstract or in a *single index term*. The last restriction should be noted in particular because sometimes relevant terms will occur only in different index term fields. For example, in Figure 1.3, neither the CAS Registry Number '7782-44-7' nor the word 'oxygen' appears in the same sentence as the word 'porphyrin' (or indeed any other of the index heading terms), so an answer of the type in Figure 1.3 would not appear in the list of candidates labeled 'closely associated' from a query 'I am interested in ''reactions of oxygen with porphyrins'' '. Accordingly, it helps if careful consideration is given to whether candidates 'closely associated' or 'anywhere in the reference' should be chosen. This is particularly important for hits in MEDLINE where each index heading is in a separate field (i.e. not 'closely associated'). As with so many situations like this, it usually is better to explore both options and to analyze answers carefully. Users of SciFinder Scholar are not restricted by search and display term costs that are a major consideration in on-line searching, and at least do not have cost factors to deter the exploration of different options in detail.

## 2.5 Working from the Reference Screen

The initial reference screen (Figure 2.4) contains links to a number of options:

- the microscope and the computer icon are used to display additional information on the answer;
- the check box is used to select specific records for further analysis;
- the **Analyze or Refine References** button is used for revision of the initial answers;

- the **Get Related** button is used to follow research through citations, to obtain substance answer sets, or to explore the search through one of three Web search engines (Chapter 4, Section 4.5); and
- the icons on the top of the screen are used for a variety of frequently required tasks (e.g. print/save).

### *2.5.1 Microscope*

The microscope is used to display the full database record and gives, for example, the type of information shown in Figures 1.3 and 1.5. There are many reasons for displaying the complete record.

First, the search terms that SciFinder Scholar has used are indicated in blue and the automatic application of truncation (Section 2.3.1), of synonyms (Section 2.3.2), and of CAS Registry Numbers (Section 2.3.4) should be checked carefully. Further, the content of the records should be noted for additional terms (particularly possible synonyms) that the user should consider adding to the query.

Second, where appropriate, the links to the CAS Registry Numbers should be tried to gain some insight into the substances indexed since the nature of the substances may give suggestions to how the question may be differently asked. For example, if substances of a particular structure are repeatedly being retrieved, the option of commencing the whole search from the point of view of chemical structures may be considered.

Finally, the Indexing Sections, Index Headings, and Supplementary Terms should be noted because not only may these give ideas as to which additional terms may be added in the query but each of these may also be used within the **Analyze References** option (Section 2.5.4).

### *2.5.2 Computer Icon*

The appearance of the computer icon means that SciFinder Scholar is directly linked to an electronic version of the full text article. SciFinder Scholar achieves this through ChemPort *http://chemport.cas.org*, which has been set up by CAS as the interface through which access to the publisher Web sites is achieved.

For example, after clicking the computer icon adjacent to the second record in Figure 2.4, the ChemPort screen (Figure 2.5) appears. The title of the article is displayed, followed by 'No Additional Fees' and 'Additional Fee' options. If the article is from a journal and the user's institution has an



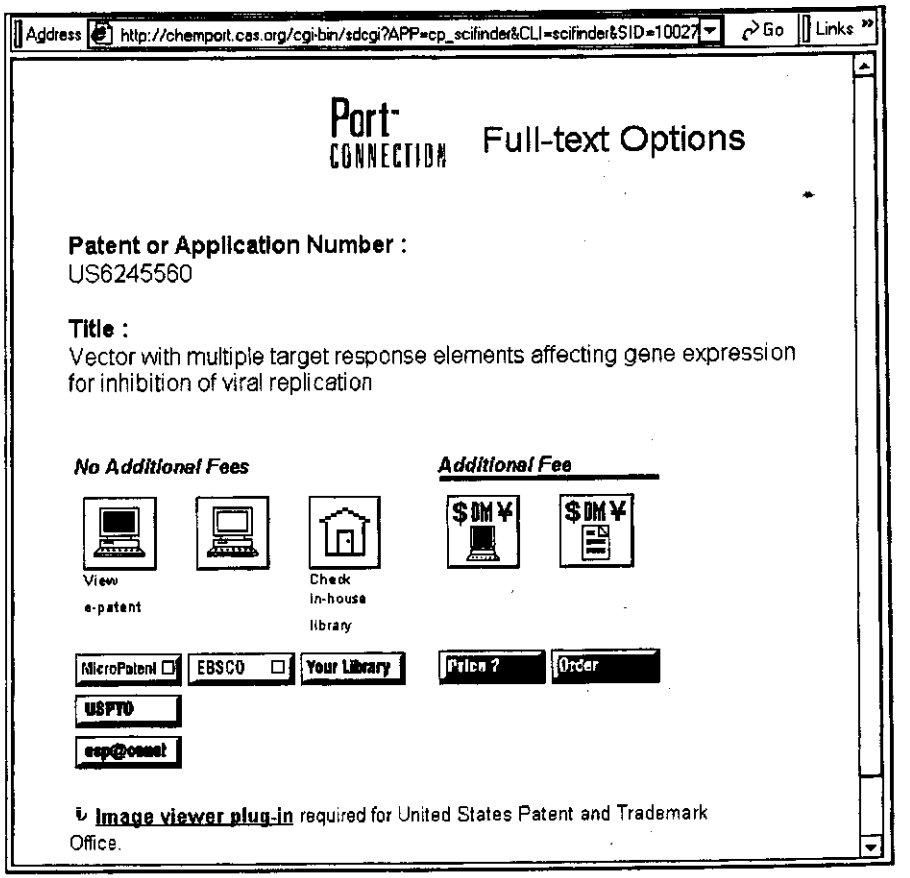



**Figure 2.5** ChemPort connection screen provides access to full text documents. ChemPort screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

electronic subscription to the e-journal, active buttons will be apparent and, when chosen, the full text will soon appear.

Most publishers now provide access to the electronic version of their journals from the mid/late 1990s onward and SciFinder Scholar provides direct links to those publishers who have agreements with CAS. The most up-to-date list of these publishers may be viewed through *http://chemport.cas.org/html/english/chemport_pub.html*.

If the record is a patent that is available electronically, then SciFinder Scholar automatically links to the full text patent either through esp@cenet, the US Patent Office or MicroPatent. To access images from the latter requires an

image viewer plug-in, which may be downloaded through a link from the ChemPort screen.

Access to the users' in-house electronic library collections, to the CAS Document Detective Service, and to other full text options is possible. The process is to click the appropriate check box at the left of the record and then to click the **Full Text** icon at the top of the screen. However, access to these options is controlled by the users' institution, so the key library coordinator should be contacted for further details. What should be remembered is that the computer icon is present only for those full text articles that are directly linked through SciFinder Scholar, and that there are a number of other options to access full text documents for the records that Scholar has identified.

The ability to link directly to full text records is a great bonus to users, who thus increasingly have immediate access to the world's scientific literature. As SciFinder Scholar links directly with more publishers, who as time passes will have increased number of articles available in electronic form, and as libraries obtain greater access to e-journals, the benefits of the integration between the primary and secondary sources will increase.

### *2.5.3 Check Box*

The check box at the left of each record is used to select specific records manually for subsequent processing, for example, to access full text documents in which direct links through SciFinder Scholar are not yet available. However, specific records are checked also for selection for refine/analyze and print/save as explained below.

The point is that although SciFinder Scholar uses very smart algorithms for the retrieval of records and although there are many precision tools and analyze/refine options, ultimately, only the user can decide exactly which answers best meet the information needs. Accordingly, some time must necessarily be spent carefully in working through actual records and checking specific boxes as needed.

### *2.5.4 Analyze and Refine References*

At the bottom of the reference screen is a button **Analyze or Refine References**. *Here are excellent tools that guide the user to narrow references in a highly predictable way.*

There are 11 options under **Analyze References** and 8 options under **Refine References** (Figure 2.6), and it will be noted that some are available under both categories. The essential difference is that **Refine References** executes instructions exactly as the user specifies, whereas **Analyze References** gives



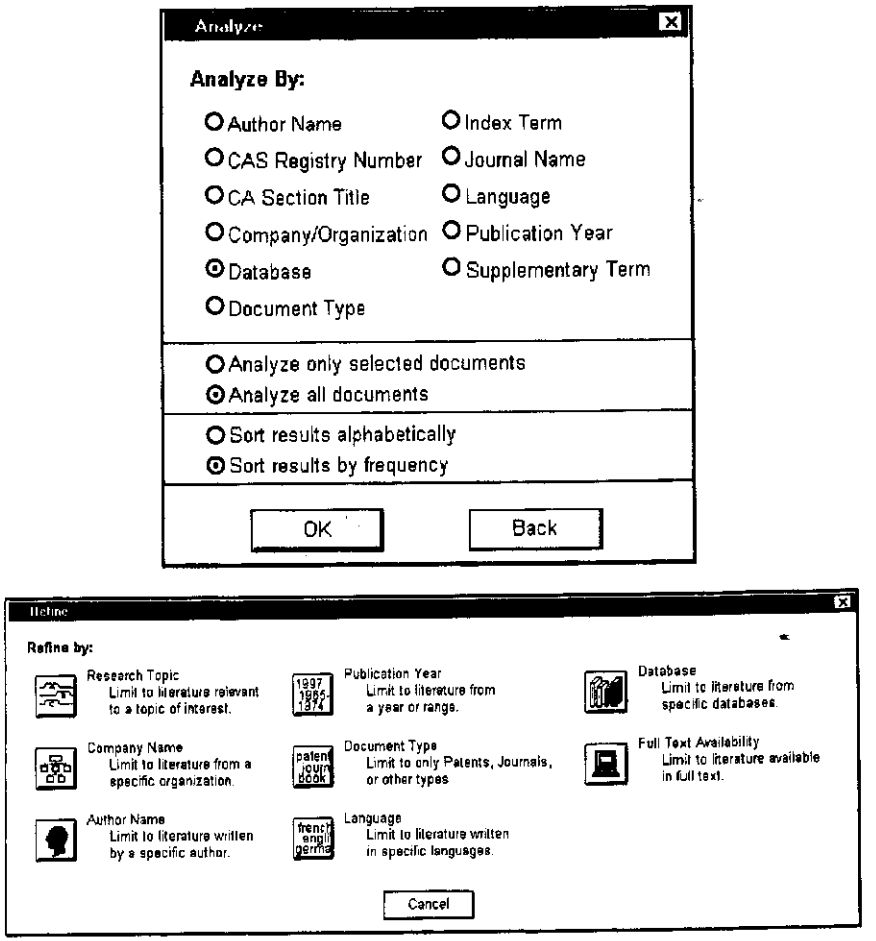



**Figure 2.6** Options under **Analyze References** and **Refine References**. SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

the user a histogram that indicates outcomes for the various paths that may be followed. Too often, users immediately jump to conclusions about what to do next, but remember all those subtle differences in records (Figures 1.3 and 1.5) for the same original document? Consider further, the even greater variation in records for different documents! Accordingly, experienced users accept the fact that *they do not know exactly* what is in the databases, and so they almost invariably take advantage of any options that guide them to alternatives. In most cases, experienced users prefer **Analyze References**.

The simplest application of **Analyze or Refine References** is the restriction of answers to one of the databases. As mentioned in the introduction to this chapter, it is advisable to search both bibliographic databases first. Even if large answer sets are obtained initially, it may be valuable to browse a few records in CAPLUS and in MEDLINE. For example, the chemist or physicist who is working on a particular technique may be surprised to see that a search on the uses of the technique retrieves records in MEDLINE. Such searches may alert the scientist to applications in a new field, and important interdisciplinary research may result.

Once the combined answer set has been obtained, analyze or refine options may be chosen to limit answers to either database, and the alternatives are shown in Figure 2.7. If the user is really sure which database to choose, then the **Refine References** option is slightly quicker; however, if the user

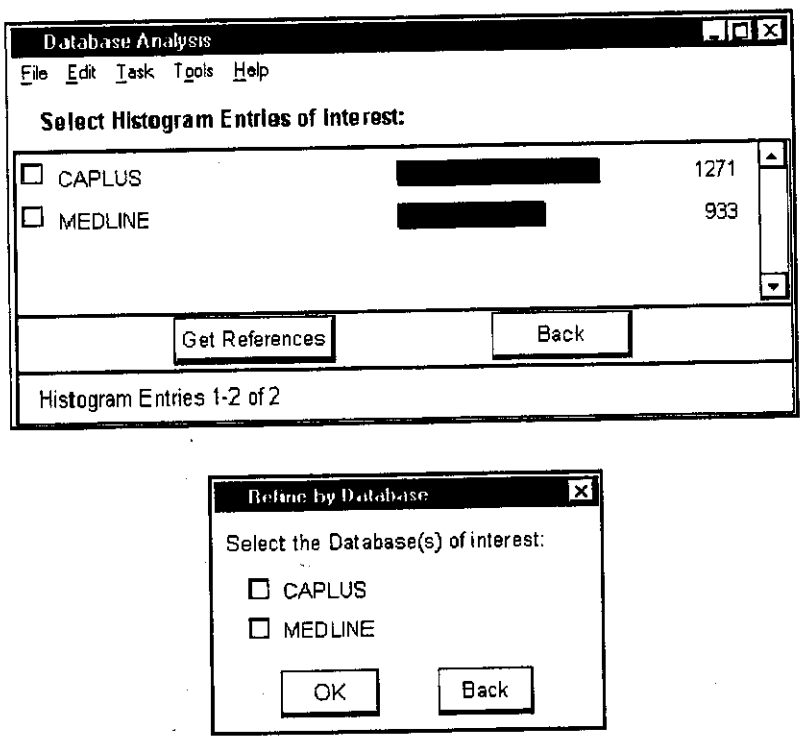



**Figure 2.7** A comparison between the operation of the **Analyze** and **Refine** options. **Analyze** gives a histogram with numbers of records in each category identified. **Refine** prompts for direct input. SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.



wants to see how many references there are in each database before making a decision, then **Analyze References** is chosen.

For the purpose of simplicity, the examples that follow in this chapter are restricted to those records from Figure 2.4 that appear in CAPLUS only. However, with equal ease, just the MEDLINE records may have been chosen or, indeed, the user may proceed from Figure 2.4 directly, in which case, subsequent analyze or refine options apply to both the databases.

While the main reason for choosing **Analyze References** is because the user is guided through alternatives, another reason is that **Analyze References** allows combinations of alternatives to be chosen. For example, consider the choices the user has if restriction of the records to patents *and* journals only is required. This may easily be achieved in a single new answer set through **Analyze References**. Thus, when **Analyze References** and then **Document Type** are chosen, a histogram (Figure 2.8) appears. Indeed, whenever options under **Analyze References** are chosen, histograms appear from which one or more options may be selected, with the added advantage of knowing the outcomes in advance. Clicking the boxes next to the top two entries, followed by **Get References**, gives the required set of answers from journals and patents.

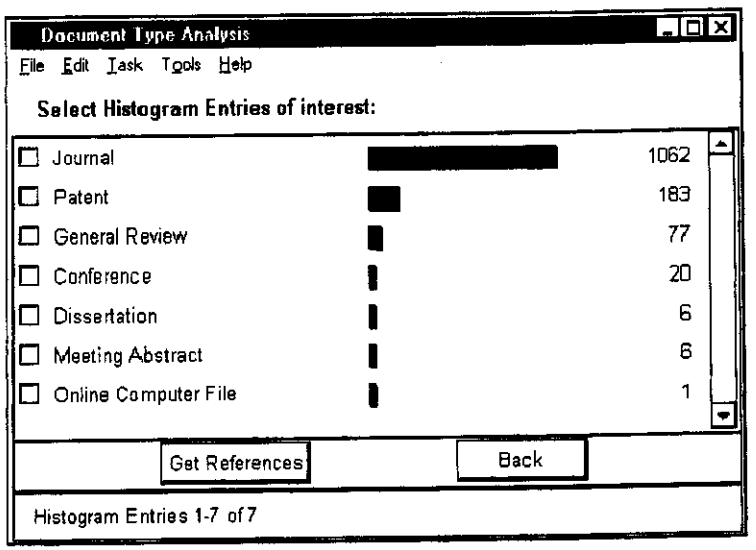



**Figure 2.8** Typical histogram obtained through **Analyze References**. (This analysis was done on the records from CAPLUS only in Figure 2.4.) SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

One of the many advantages of SciFinder Scholar is that it guides the searcher. Recall that once a query under **Explore by Research Topic** is entered, SciFinder Scholar suggests the candidate options and indicates the number of records within them. In a similar way, **Analyze References** provides options and indicates outcomes in advance.

While the result of the operation in Figure 2.8 is that various document types may be selected, the same processes are available through the other **Analyze References** options (Figure 2.6). Many important possibilities are noteworthy.

### Analyze by Author and Company

First, **Analyze References** by **Author Name**, and then display of the histogram in alphabetical order may alert the user to issues not previously considered. For example, when the references in Figure 2.4 are analyzed by the author, some of the entries for Erik De Clercq are shown in Figure 2.9. However, careful examination of other entries in the complete histogram reveals entries for 'Clercq E De', 'Clercq Erik De', 'de Klerk E', and 'Declercq E', and these would not have been retrieved through **Refine References** by **Author Name** unless refinements specifically related

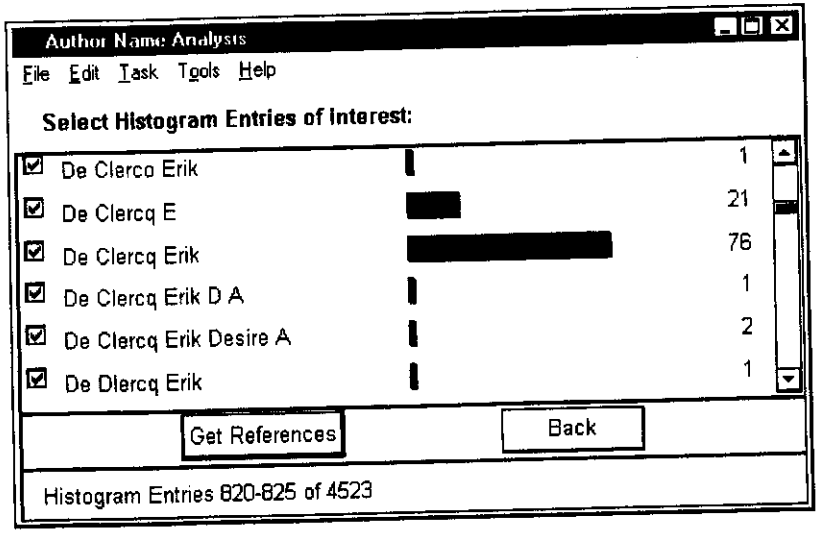



**Figure 2.9** Some entries for author Erik De Clercq identified under **Analyze References** by **Author Name** from records in CAPLUS (Figure 2.4). SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.



to these alternative spellings were considered. The issue here is mainly that authors, and the editors of the journals in which they publish, may vary in the way the names are entered, and the **Analyze References** by **Author Name** provides a very good method to handle such variations. Further issues relating to searching for authors are discussed in Chapter 4, Section 4.2.

Second, **Analyze References** by **Company/Organization** provides a histogram of listed affiliations that helps the user through the extremely complicated issues involved with searching for names of institutions (Chapter 4, Section 4.4). For example, the major affiliation listed in the references (Figure 2.4) is the Rega Institute for Medical Research, Belgium, and there are at least 26 separate listings for this institution in this answer set (Figure 2.10). The multiplicity of entries in this case is due to the different ways the authors listed their affiliation in the original document. Through the **Analyze References** by **Company/Organization**, SciFinder Scholar is presenting the full list to users who need to carefully go through the histogram to pick out all the relevant entries.

The important point relating to queries involving author and company terms is that searching databases may be very complicated. There may be all types of pitfalls for the unwary searcher! However, **Analyze References** allows the searcher to instantly recognize issues, and the ability of SciFinder Scholar to guide the searcher in this way must never be underestimated.

### Analyze by Index Term

Third, **Analyze References** by **Index Term** provides a histogram (Figure 2.11) of index headings, which are discussed in Chapter 1, Sections 1.5 and 1.6. **Analyze References** by **Index Term** (and also by **Supplementary Term**) has two main uses. First, the choice of appropriate entries enables the user to obtain more precise answer sets through the use of the *systematic terms and keywords* applied to the records by indexers (e.g. choice of 'Virucides and Virustats' and 'Antiviral agents' focuses on records with the index terms). Second, index terms usually reflect important terminology in the field and may alert the user to terms that may need to be included in the original question. Indeed, another way to ask for information on the inhibition of HIV replication may well be to enter – 'I am interested in ''virucides (virustats, antiviral agents) for HIV'' '.[3]

[3] A similar analysis of index headings in MEDLINE indicates the extensive use of the heading ANTI-AIDS, which may also be appropriate.

LAB. CHEMOATHERAPY, REGA INST. MED. RES., LOUVAIN, B-3000, BELG.
LAB. PHARM. CHEM., REGA INST. MED. RES., LOUVAIN, BELG.
LABORATORY OF EXPERIMENTAL CHEMOTHERAPY, REGA INSTITUTE FOR MEDICAL RESEARCH, LOUVAIN, B-3000, BELG.
REGA INST. FOR MED. RESEARCH, KATHOLIEKE UNIVERSITEIT LEUVEN, LOUVAIN, BELG.
REGA INST. FOR MEDICAL RES., KATHOLIEKE UNIV. LEUVEN, LOUVAIN, B-3000, BELG.
REGA INST. MED. RES., CATHOL.UNIV. LEUVEN, LOUVAIN, B-3000, BELG.
REGA INST. MED. RES., KATHOL UNIV. LEUVEN, LOUVAIN, B-3000, BELG.
REGA INST. MED. RES., KATHOL. UNIV. LEUVEN, LOUVAIN, B-3000, BELG.
REGA INST. MED. RES., KATHOL. UNIV. LEUVEN, LOUVAIN, BELG.
REGA INST. MED. RES., KATHOL. UNIV. LOUVAIN, LOUVAIN, B-3000, BELG.
REGA INST. MED. RES., KATHOL. UNIV., LEUVEN, LOUVAIN, B-3000, BELG.
REGA INST. MED. RES., KATHOL. UNIV., LOUVAIN, B-3000, BELG.
REGA INST. MED. RES., KATHOLIEKE UNIV. LEUVEN, LEUVEN, B-3000, BELG.
REGA INST. MED. RES., KATHOLIEKE UNIV. LEUVEN, LOUVAIN, B-3000, BELG.
REGA INST. MED. RES., KATHOLIEKE UNIV. LEUVEN, LOUVAIN, BELG.
REGA INST. MED. RES., KATHOLIEKE UNIV. LEUVEN, LOUVAIN, 3000, BELG.
REGA INST. MEDICAL RES., KATHOLIEKE UNIV. LEUVEN, LEUVEN, B-3000, BELG.
REGA INST. MEDICAL RES., KATHOLIEKE UNIV. LEUVEN, LOUVAIN, B-3000, BELG.
REGA INSTITUTE FOR MEDICAL RESEARCH (P.C., J.A.E., ED.C., ZD.) AND CENTER FOR HUMAN GENETICS (G.R.,R.S., G.D.)
REGA INSTITUTE FOR MEDICAL RESEARCH, KATHOLIEKE UNIVERSITEIT LEUVEN, LOUVAIN, B-3000, BELG.
REGA INSTITUTE FOR MEDICAL RESEARCH, KATHOLIEKE UNIVERSITEIT LEUVEN, LOUVAIN, BELG.
REGA INSTITUTE MEDICAL RESEARCH, KATHOLIEKE UNIV. LEUVEN, LOUVAIN, B-3000, BELG.
REGA INSTITUTE MEDICAL RESEARCH, KATHOLIEKE UNIVERSITEIT LEUVEN, BRUSSELS, BELG.
REGA INSTITUTE MEDICAL RESEARCH, KATHOLIEKE UNIVERSITEIT LEUVEN, LOUVAIN, B-3000, BELG.
STICHTING REGA V. Z. W., BELG.
STICHTING REGA V.Z.W., BELG.

**Figure 2.10** Entries for the Rega Institute from CAPLUS records in the answer set (Figure 2.4).

### Other Analyze Options

Similarly, any of the other options under **Analyze References** in Figure 2.6 may be chosen. These histograms have many uses and a summary of uses for all analyze options is given in Table 2.2. **Analyze** is an extremely important tool and the challenge for the user is to employ it creatively. Often it is simply a matter of knowing how it works, of observing outcomes from some alternatives, and of thinking how particular options will give the user insights into the literature. Appendix 4 gives outcomes of other analyze options. Think about them, and try them!



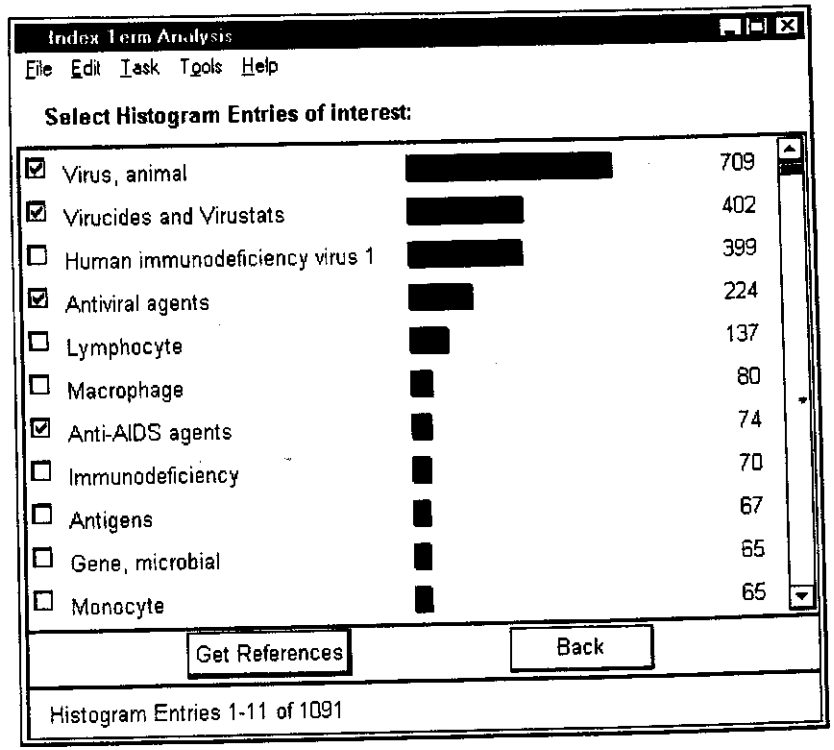



**Figure 2.11** Histogram of CAS Index Terms. This result was obtained by restricting answers in Figure 2.4 to CAPLUS only. A histogram of MEDLINE Index Terms may be obtained in a similar way (see Figure 5.1). SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

### Uses of Refine References

While **Analyze References** provides many advantages over **Refine References**, the latter provides one unique option, namely, **Refine by Research Topic**, in which the terms in the title, abstract, and index fields are searched. This option is particularly important when it is necessary to restrict answers by inclusion of terms that are not part of indexing, and hence that occur in titles or abstracts only. For example, one important property of polymers is the ratio of molecular weight to molecular number (Mw:Mn), but information on this very specific technical issue is likely to occur only in the abstract text. Accordingly, if information on this ratio for linear low-density polyethylenes is required, an **Explore by Research Topic** 'I am interested in ''linear

**Table 2.2** Some uses of **Analyze** options

| Analyze option | Uses | Notes |
|---|---|---|
| **Author Name** | A single author may be represented in a number of different ways (e.g. compare Figures 1.3 and 1.5); so, if all publications from an author are required, then **Analyze References** by **Author Name** followed by **Sort** results alphabetically should be used | For further information on searching for authors, see Chapter 4, Section 4.2 |
| **CAS Registry Number** | Sort results by frequency gives a histogram of all the CAS Registry Numbers in the records. Useful when interested in the type of substances involved. Indexing of substances in MEDLINE is more generic, so **Analyze** by **CAS Registry Number** mostly is done through answer sets from the separate databases | To find substances in traditional Chinese medicines, try **Explore by Research Topic** 'traditional Chinese medicines', followed by **Analyze References** by **CAS Registry Number**. Many substances that are not medicines will be retrieved, but the list may be examined manually to identify key medicines |
| **CA Section Title** | (Applies only to records from CAPLUS.) Useful to restrict answers to broad categories. **Analyze References** by **CAS Section Title** usually gives histograms of less than 100 entries, so the list to review is relatively small–and different areas may be readily identified | **Explore by Research Topic** 'I am interested in ''aids'' ' will retrieve hits for the disease AIDS, and also audio-visual aids, filtering aids, etc. While additional concept terms or **Analyze References** by **Index Term** may be used, **Analyze References** by **CAS Section Title** is another (and quick) option |
| **Company/ Organization** | Companies/universities/university schools and departments may be listed in a very large number of ways; the analyze option offers considerable help in identifying relevant terms | For further information on searching for company names, see Chapter 4, Section 4.4 |
| **Database** | Used for identifying numbers of records in each database (CAPLUS/MEDLINE), and then for obtaining records in just one of the databases (if required) | |



**Table 2.2** (*continued*)

| Analyze option | Uses | Notes |
|---|---|---|
| **Document Type** | Different types (reviews, clinical trials etc.) may be identified and may help in narrowing answers to meet precise requirements | CAPLUS and MEDLINE have approximately 20 and 40 document types, respectively |
| **Index Term** | Very useful for precision searches based on systematic terminology and for identification of search terms (synonyms) to assist comprehensive retrieval | The search Figure 2.2 produces answer sets that may be analyzed by Index Term (Figures 2.11 and 5.1). Terms such as virustats, antiviral agents, antiaids, etc. are displayed and may be considered as possible search terms in new searches |
| **Journal Name** | Used to identify key journals, e.g. those that should be held in the institution's collection or those that cover a particular topic (perhaps to decide on where to submit a manuscript for publication) | |
| **Language** | Used to restrict answers to one (or more) languages in the original publication | |
| **Publication Year** | Used to determine trends in reporting on the topic, and to obtain acceptably small answer sets to download, or process further | Use **Sort** by **Alphabetical Order** to see list of years in chronological order |
| **Supplementary Term** | Similar to Index Term, but Supplementary Terms are not part of systematic vocabulary | |

*Note*: **Analyze** operates on any initial set of references, and these comments are restricted to its use to narrow an answer set obtained from an **Explore by Research Topic**. Depending on requirements, answers may be sorted by frequency or alphabetically.

low density with polyethylene or lldpe'' ' is first conducted and the initial answers (>7200) are then refined by research topic 'mw mn'. This gives around 250 answers, nearly all of which contain actual data for the Mw:Mn ratio.

There are two other important points to note. First, like all other refine and analyze options, **Refine by Research Topic** operates at the 'and' (or 'not') level and it is not possible to **Refine by Research Topic** in such a way that the terms now entered under refine are 'closely associated' with other terms

previously searched. The implication is that some level of precision may be lost. Thus, if two concepts (A, B) are entered at the Explore level and the candidate in which they are 'closely associated' is chosen, and if the answers are refined with a third concept (C), then the outcome would effectively be an answer set: (A closely associated with B) AND C. On the other hand, entry of three concepts (A, B, C) at the Explore level would give, *inter alia*, a candidate where all three are closely associated. Nevertheless, as mentioned in Section 2.2, entering several concepts at the Explore level may produce a very lengthy list of candidates, so the right balance needs to be achieved for the special problem being solved.

Actually, the inability to closely associate terms under **Refine by Research Topic** is a restriction that applies to this option only. All other refine and analyze options necessarily have to be performed at the 'and' level because they involve separate fields for the data being searched.

Second, if alternative terms need to be entered through **Refine by Research Topic**, they must be connected with 'or' because the option, which is available at the Explore level (Section 2.3.2), of placing alternatives in parentheses is interpreted under Refine by Boolean 'and'. For example, **Refine by Research Topic**: 'men or women' is acceptable, but **Refine by Research Topic** 'men (women)' is interpreted as requiring both concepts to be present, and this may not be the intention.

## 2.6 Saving and Processing Answer Sets

Any reference set or individual reference may be printed or saved in a number of formats for future reference. Instructions on how to proceed are available through a variety of Help messages. When answers are saved, a warning appears: 'Use of saved information is subject to copyright and data use restrictions as specified in the SciFinder end user license agreement'. It is relevant to point out that users need to accept the License Agreement every time they enter SciFinder Scholar, and that Clause 4 mentions the restrictions relating to saved answers.

Actually, SciFinder Scholar users currently are restricted to saving less than 100 records at a time. If slightly more than 100 records are required, and if it is not possible to analyze or refine to obtain smaller answer sets without the loss of relevant records, then an option is to **Analyze References** by **Publication Year** alphabetically and finally, to check boxes so that an answer set of less than 100 records is obtained. This set may be saved and the process repeated so that all the required answers are saved. Having



mentioned that, it is imperative that usage restrictions are observed and if in any doubt, users should check with their key coordinator on procedures. The CAS has always provided generous access to academic institutions at highly discounted prices, and it is essential that staff and students cooperate fully with their key coordinator to ensure that no usage of SciFinder Scholar jeopardizes these arrangements.

## 2.7 Applying Scientific Method to Information Retrieval

This chapter has described the fundamentals of using **Explore by Research Topic** and has outlined many of the ways in which SciFinder Scholar works behind the scenes to guide the user. However, it helps if the searcher applies 'scientific method' to information retrieval (Figure 1.1), and in particular, to work through Steps 3 to 7:

Step 3 conceptualization of initial search query;
Step 4 initial search;
Step 5 careful examination of initial answer sets;
Step 6 revision of search query based on observation of initial answers; and
Step 7 exploration of alternative search options.

To illustrate these steps, consider a question to find information on the ways to determine the mass of quarks (a group of subatomic particles). The conceptualization of the query requires an understanding that there are essentially three concepts (determination, mass, quarks). With the knowledge that SciFinder Scholar will best guide the searcher if prepositions are entered in the query, the initial search may be 'I am interested in ''the determination of the mass of quarks'' '.

The list of candidates indicates that there are more than 7 million records with the concept '**determination**', almost 900 000 records with the concept '**mass**', and more than 33 000 with the concept '**quarks**'. Presently, all the concepts are closely associated in just over 260 records. It is also noted that the three concepts appear anywhere in the record in almost 3000 records and that there are more than 5500 records with the concepts '**mass**' and '**quarks**' closely associated.

The user thus makes a mental note that many of the records with '**mass**' and '**quarks**' closely associated will probably be of interest, but before going down this path, it helps if the user checks what is being searched. So, careful examination of the 260 records in which all the concepts are closely associated is first undertaken, since it usually is quickest to examine

the process of the search through the most specific answers. It immediately becomes apparent that the concept '**determination**' has included synonyms on the basis of 'analysis', and from observation of some of the records, it is apparent that the answers are not precise enough.

Accordingly, the question is reiterated 'I am interested in ''measurement of the mass of quarks'' ' and SciFinder Scholar indicates that there are 79 records with all the three concepts closely associated. These certainly are looking more like what is required, and indeed a review by Richard Partridge in 1999 titled 'Heavy quark production and decay: t, b, and onia' even mentions in the abstract 'the top quark mass has been measured with an uncertainty of <3 %'.

Just as in any scientific experiment, the scientist keenly observes what is happening and immediately notes that 'quark' in the above abstract is not highlighted. Yet, 'top quark mass' clearly is an important phrase! Further, the indexing mentions 'top quark', 'bottom quark', and 'charm quark', and neither the word 'quark' nor any of the CAS Registry Numbers for these three particles (yes, it is noted they each have CAS Registry Numbers!) are highlighted. Does this mean that the SciFinder Scholar algorithm has broken down?

The point is that any artificial intelligence operates within smaller boundaries than human intelligence; so, although SciFinder Scholar automatically handles many issues, it may not cover all the possibilities. Scientific experiments do not always obey initial instructions! In the 79 records, it is noted that although 'mass', 'masses', and various words with the word stem 'measur. . .' (i.e. measuring, measurement, measurements, etc.) are searched, in this case, the singular 'quark' is not included as a synonym.[4]

Accordingly, the experiment is repeated: 'I am interested in ''measurement of the mass of quarks (quark)'' '. Now, one candidate in the list indicates 546 records with '**measurement**' and '**mass**' and either '**quarks**' or '**quark**' closely associated, while another in the list indicates 545 with '**measurement**' and '**mass**' and '**quark**' closely associated. The algorithm clearly has allowed for the plural 'quarks' when the singular 'quark' was entered. One of the 546 records is titled 'Top quark results from the Tevatron' by Serban Protopopescu and mentions in the abstract a 'top quark mass of 174.1 ± 5.4 GeV/c2'. The careful examination of initial answer sets (Step 5) and the revision of search query (Step 6) has clearly paid off and a reasonably precise and comprehensive answer set has been obtained.

[4] At least currently! The effectiveness of the algorithm is constantly being monitored, and improvements are implemented as appropriate from time to time. Nevertheless, at the time it was searched, this example illustrates why it is helpful if the searcher carefully examines the answers.



From this answer set, many alternatives may be followed. For example, to focus on the strange quark, **Refine References** may be chosen to allow for words in titles and abstracts as well as indexing. The entry here may be: '183748-09-6 or strange quark or s quark', as from some of the 546 records, it is observed that the CAS Registry Number for strange quark is 183748-09-6 and that 's quark' is commonly used.

Alternatively, some options under **Analyze References** may be chosen and in the case of analyze by **Index Term**, the major entries are indicated in Figure 2.12.

Interestingly, when **Analyze References** by **Journal Name** is chosen, it is found that about 10 % of the records are preprints from the Los Alamos National Laboratory Preprint service and the exact url appears. Clicking on this url provides immediate access to the electronic version of the documents.

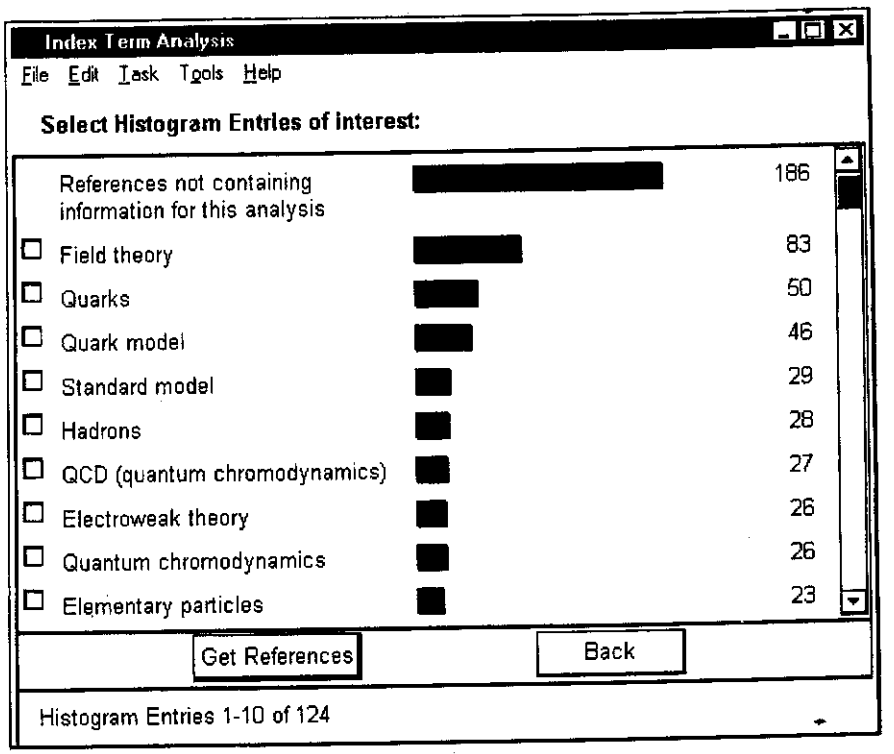



**Figure 2.12** Most-listed Index Terms in CAPLUS for search on measurement of mass of quark(s). SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

Depending on what the user has found, other search options could be explored. In particular, it would now be apparent that all the individual quarks have different CAS Registry Numbers and the search could commence from any one of a number of options.

## 2.8 Summary

Central to the operation of SciFinder Scholar is the option to **Explore by Research Topic** using simple natural language questions. There are a number of issues that need to be considered in the statement of the initial query and these are summarized in Appendix 3. In general, it is advisable to include prepositions in the query because more candidates are presented, although if too many concepts are entered, the list of candidates may be too lengthy.

Users must look at the list of candidates to see how SciFinder Scholar has interpreted the query and must think carefully about a number of issues in choosing the most appropriate set of candidates. In general, SciFinder Scholar identifies different concepts on the basis of the presence of the prepositions and conjunctions in the query, displays candidates where the concepts are within the same sentence ('closely associated') or simply within the record ('anywhere in the reference'), displays candidates with combinations of the individual concepts, and displays the individual concepts.

Individual answers should be checked for hit terms, and modifications should be made particularly where additional synonyms become apparent.

**Analyze References** provides many powerful options for the revision of answer sets, including the important option to revise by the systematic terms entered by indexers, while **Get Related** offers further powerful options (Chapter 4, Section 4.5).

However, it must be remembered that SciFinder Scholar is the front end to databases of many millions of records that are primarily made up of records with text from authors and indexers. When a search is being constructed, it helps if the user considers what type of terms an author and an indexer may have entered. In particular, it should be remembered that authors do not follow any universal policies in writing up text in titles and abstracts and that there may be very considerable variation in the way different authors write about their work.

Having said that, SciFinder Scholar guides the searcher through the investigation. First, the searcher is guided by a list of candidates and second, the searcher is guided through **Analyze References**. In this way, SciFinder Scholar helps the user overcome many of the issues that arise because of the size and complexity of the databases.



## 2.9 Exercises

### *2.9.1 General Questions on the Operation of SciFinder Scholar*

2.1 Which of the categories are common to **Analyze References** and **Refine References** and which are different (Figure 2.6)? When categories are in common, what criteria are used to decide which one to choose?

2.2 What parts of records are searched in **Explore by Research Topic** ?

2.3 Suggest reasons why SciFinder Scholar does not allow the user to search for a topic and the name of an author *at the same time*?

2.4 How many concepts will be identified in each of the following questions:

2.4.1 Antibiotic resistance in pigs or poultry;

2.4.2 Antibiotic resistance in pigs and poultry;

2.4.3 Resistance to antibiotics in pigs or poultry;

2.4.4 Resistance to antibiotics in pigs (poultry);

2.4.5 Pig antibiotic resistance.

2.5 In which ways will the nature of the final answer sets differ from the following three approaches (i.e. how do the searches differ)?

- **Explore by Research Topic**: I am interested in 'resistance to antibiotics in pigs or poultry' and choice of one of the candidates that mentions all four concepts.
- **Explore by Research Topic**: I am interested in 'resistance to antibiotics', choice of the candidate that has both concepts 'closely associated', followed by **Refine References** with **Research Topic** 'pigs or poultry'.
- **Explore by Research Topic**: I am interested in 'resistance to antibiotics', choice of the candidate that has both concepts 'closely associated', followed by **Refine References** with **Research Topic** 'pigs (poultry)'.

### *2.9.2 On-Line Questions*

2.6 Find information on:

2.6.1 Scifinder or Scifinder Scholar;

2.6.2 antibiotic resistance in pigs or poultry;

2.6.3 forensic studies on drug overdose;

2.6.4 methods for prediction (warning) of earthquakes;
2.6.5 pharmaceuticals for the treatment of glaucoma;
2.6.6 defense pheromones used by insects;
2.6.7 studies on the DNA of cheetahs (acinonyx);
2.6.8 conducting polymers in textiles;
2.6.9 materials for the replacement of bones (e.g. of knees or hips);
2.6.10 natural products (particularly fatty acids) from Shewanella;
2.6.11 removal of free fatty acids from edible oils;
2.6.12 removal of the bitter components from the juice of citrus fruits;
2.6.13 theoretical calculations on the entropy of DNA;
2.6.14 inhibitors for tyrosine kinase;
2.6.15 bile acids that are generally regarded as safe (GRAS);
2.6.16 biological availability, pharmacokinetics, and absorption of itraconazole;
2.6.17 age of onset of coeliac (celiac?) disease or disorder;
2.6.18 the *Pteropus* species (which are also known as fruit bats or flying foxes);
2.6.19 why fish do not freeze in the cold water of the Antarctic or Arctic oceans;
2.6.20 the molecular weight to molecular number ratio of linear low-density polyethylenes.

# Chapter 3
# Explore by Chemical Substance

## 3.1 Introduction

There is hardly any scientific discipline that does not embrace chemicals! Chemicals may be relatively simple like the individual atoms or may be complex like a single DNA molecule that has many billions of atoms. Creating a database for the vast numbers and varieties of chemicals is a very major challenge!

However, the issues go further than that. Sometimes, the bonds that hold atoms together are difficult to describe. Sometimes, a single substance is referred to by many different names. Sometimes, substances are incompletely described in the literature. Sometimes, substances are loosely described, for example, the agriculturist may refer to N:K:P ratios in soils and is talking about something very different from what a chemist understands as N, K, P (i.e. the individual atoms, not the ions, present in salts).

Descriptions of molecular structures occupy large sections of university chemistry courses. Remember the lectures: ionic and covalent bonding, valence bond and molecular orbital theory, resonance and tautomerism, coordination compounds, molecular associations, the structures of metals and alloys, the formation and structures of polymers? Tied in with all this are the issues of nomenclature. Arguably, there is not a more complex area in science and consequently, there are many issues for those who build comprehensive substance databases to consider!

Over the years, the indexing staff at Chemical Abstracts Service (CAS) have addressed all the issues involved with the description of substances and, working with scientists worldwide, continue to address new issues as they arise. For example, combinatorial chemistry and supramolecular chemistry

are two relatively recent developments that challenge any electronic indexing system for substances. The aim is to produce as simple and as systematic a description of substances as possible, and then to index substances in such a way so as to facilitate comprehensive and precise retrieval of information.

This is achieved through the entry of substances in a chemical substance database in which each unique substance is given a single registration number. The so-called CAS Registry Numbers are now used universally to identify substances, and the complete CAS chemical substance database that contains all the CAS Registry Numbers is available through SciFinder Scholar.

The CAS Registry Numbers may be found in REGISTRY through chemical structure, molecular formula, or chemical name–based search terms. These choices help overcome the enormous problems associated with the names of substances, and once the CAS Registry Numbers have been identified, any search based on them will help overcome the problems associated with the use of many different names for the same substance.

This chapter briefly describes key aspects of the Registry System, then explains how to find CAS Registry Numbers, and finally comments on issues relating to searching for information on substances in SciFinder Scholar.

## 3.2 Registration of Substances

The CAS Registry System uses the valence bond theory for atoms, so structure representations mostly are the same as those through which scientists normally communicate. However, simple valence bond theory may be inadequate or may be interpreted differently by scientists in four instances: resonance, tautomerism, $\sigma$-bonding, and $\pi$-bonding. Valency bond theory also has limitations in classes of substances *inter alia* polymers, cyclophanes, and other supramolecular assemblies.

Thus, sometimes CAS needs to apply special policies for the representations of structures. They should not be considered 'quirky' but instead should be considered as a necessary part of the process in building a comprehensive *electronic* database for substances.

### *3.2.1 CAS Registry Numbers*

The CAS Registry Numbers are unique descriptors for chemical substances. They are assigned in chronological order, and so the numbers have no chemical significance. The complete source of Registry Numbers is the Registry database (Chapter 1, Section 1.7).

A CAS Registry Number is given to *each unique substance*, so a single amino acid variant in two proteins will mean two different registrations.



Similarly, the sodium and potassium salts of a carboxylic acid will have different CAS Registry Numbers, and these in turn will be different from the parent acid. However, in some cases, different forms or modifications of a substance will have the same CAS Registry Number, and, for example, the CAS Registry Number for a polymer is based on the starting components alone; the method of polymerization (conditions, catalysts, etc.) is not relevant.

The CAS Registry Numbers are used systematically in CAS databases as the reference and indexing points for substances (for a summary of the policies, see Table 3.1), so searches on specific substances in CAS databases nearly always should involve searches on CAS Registry Numbers. The importance of this has already been noted (Chapter 1, Section 1.5.3).

On occasions, substances that have CAS Registry Numbers will not have any literature citations. There are a number of reasons for this, although it occurs most commonly because organizations may apply to CAS for CAS Registry Number registrations even though they have not published data on the substances (Appendix 5.5.3).

### *3.2.2 Policies for Substance Descriptions*

The applications of many of these policies are best illustrated through examples, and a number of categories are overviewed in Table 3.1.

Built into SciFinder Scholar are extensive algorithms that automatically allow for the interpretation of many of these policies, and which automatically handle issues of resonance, tautomerism, and substances that may be represented in different ways (e.g. open and ring forms of carbohydrates, pentavalent phosphorous halides, etc.). However, it pays for the scientist to consider why particular answers have been retrieved and if in doubt about what is occurring, to seek explanations. In the sciences, the keen *inquirer and observer* makes significant discoveries, and this is also true with information retrieval.

## 3.3 Searching for Substances: The Alternatives

Currently, there are four main ways to search for information on substances in SciFinder Scholar. Perhaps, the easiest way for those not reasonably familiar with the complexities of chemical substances is simply to search under **Explore by Research Topic**. For example, it was seen in Chapter 2, Section 2.7 that careful analysis of the results from a search on a particular group of substances (quarks) revealed relevant CAS Registry Numbers.

**Table 3.1** Overview of principal indexing issues for substances in REGISTRY

| Issue | Summary of general indexing | Example |
|---|---|---|
| $\sigma$-bonds (where both bond electrons are provided by one of the atoms in the bond) | Represented as a double bond between the atoms | Appendix 5.1.4 |
| $\pi$-bonds | Represented as a single bond between all the participating atoms | Appendix 5.3.4 |
| Stereoisomers | Represented by stereochemical descriptors in the name field and in the structure | Appendix 5.1.3 |
| Isotopes | Hydrogen isotopes represented in the formula field; all isotopes represented in the name and structure fields | Appendix 5.1.2 |
| Resonance | Special bond description ('normalized bond') | Appendix 6 and Section 3.4.2 |
| Tautomerism | Individual valence bond structures of tautomers as described in document are entered | Appendix 6 and Section 3.4.2 |
| Alloys | When precise ratios of elements are known, the ratios are listed in the name and composition fields | Appendix 5.2.2 |
| Salts | Generally indexed as two component substances with the acid and the base as separate components | Appendix 5.2.1 |
| Mixtures, hydrates, host–guest complexes | Indexed as substances containing separate components | Appendix 5.2.3 |
| Metal complexes | A variety of registrations depending on nature of the complex | Appendix 5.3 |
| Polymers | A variety of registrations but mainly as substances containing separate components for the monomers | Appendix 5.4 and Section 5.5 |
| Peptides/proteins | Peptides have sequence data; those with less than 50 residues also have structure data | Appendix 5.4.3 and Section 5.4 |
| Nucleic acids | Nucleic acids have sequence data; those with less than 5 residues also have structure data | Appendix 5.4.4 and Section 5.4 |
| Incompletely defined substances | A variety of registrations depending on nature of the substance | Appendix 5.5.1 |
| Minerals | Compositions, where known, are given in the name and composition fields; otherwise, indexed by name only | Appendix 5.5.2 |
| Natural oils, fats, etc. | Described by common or trade name, or by source in name and definition fields; in general, they should be searched in CAPLUS by CA Index Name rather than by CAS Registry Number | Appendix 5.5.3 |



Searches based on names of substances and on the CAS Registry Numbers thus identified may then be conducted.

All the other three ways start with **Explore by Chemical Substance** and use either structure, name, or molecular formula – based terms (Figure 3.1), and these are discussed in this chapter. Once the substances are found, the user clicks **Get References** and follows options to the answers in CAPLUS and MEDLINE. *The search terms used for the substances in this crossover from REGISTRY are CAS Registry Numbers only*, but as noted in Figure 1.3, this is exactly what is required!

*Searching by name* usually requires an exact match of the entry with a complete name of a substance in the database. It is the method of choice if a common or trivial name is known, and substances *inter alia* cholesterol, penicillin G, morphine, sodium acetate, and acetylene are easily found by this method. Indeed, most of the substances in biology, engineering, physics, the material

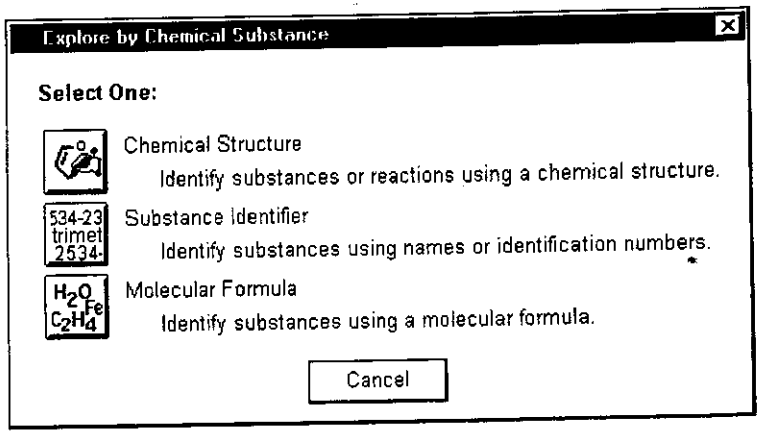
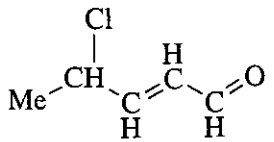



**Figure 3.1** Options for **Explore by Chemical Substance**. SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

sciences, and the pharmaceutical and medical sciences have relatively simple names, and scientists working in these fields will usually be successful in finding the substances they commonly use through a name search.

However, it may be a challenge to find by name the specialist substances used in research chemical, biochemical, and pharmacological laboratories. Even simple substances like (1) may be tricky to find by name-based terms,[1] and in any case, the scientist would need to consider how the four possible stereoisomers of (1) would be named. Probably, a structure search would be the easiest solution in this case!




(1)

*Searching by molecular formula* may require some knowledge of how formulas are entered, but even scientists with only very basic knowledge of substances may easily calculate formulas. However, many substances may share the same formula, and in the case of a search on the molecular formula for (1) (C5 H7 Cl O), around 140 substances are found. These may take a while to look through. Probably, the searcher would find it easiest and quickest to look through these answers by a search refinement based on chemical structure; so again, knowledge of structure searching is valuable!

*Searching by structure* may require some knowledge of how substances are represented in the database, but once a few rules are understood, then in the majority of cases, structure searches will be the method of choice. Structure searches also open up possibilities to find substances related by structure, and this often is a very important aspect that is not possible through name or formula searches.

In the case of substance (1), it is a simple matter in SciFinder Scholar to draw and perform an exact search on the structure. All stereoisomers and isotopically substituted substances are retrieved and the answer list may easily be examined for the most relevant answers.

[1] If 4-chloro-2-pentenal is searched as a name, then SciFinder Scholar finds 20 'possibilities' and in fact one of them is correct.



## 3.4 Explore by Chemical Structure

Thus, of the three options under **Explore by Chemical Substance** (Figure 3.1), the most commonly used is **Explore by Chemical Structure**. When **Chemical Structure** is selected, the structure drawing screen (Figure 3.2) appears and the structure is drawn in a manner similar to that used in most computer chemical structure drawing programs.

After the structure is drawn, the user clicks **Get Substances** and chooses one of the options in Figure 3.3. But what do they mean? When one of the options is chosen, the structure search is performed and answers (i.e. substances from REGISTRY) are obtained. But what type of answers are produced?

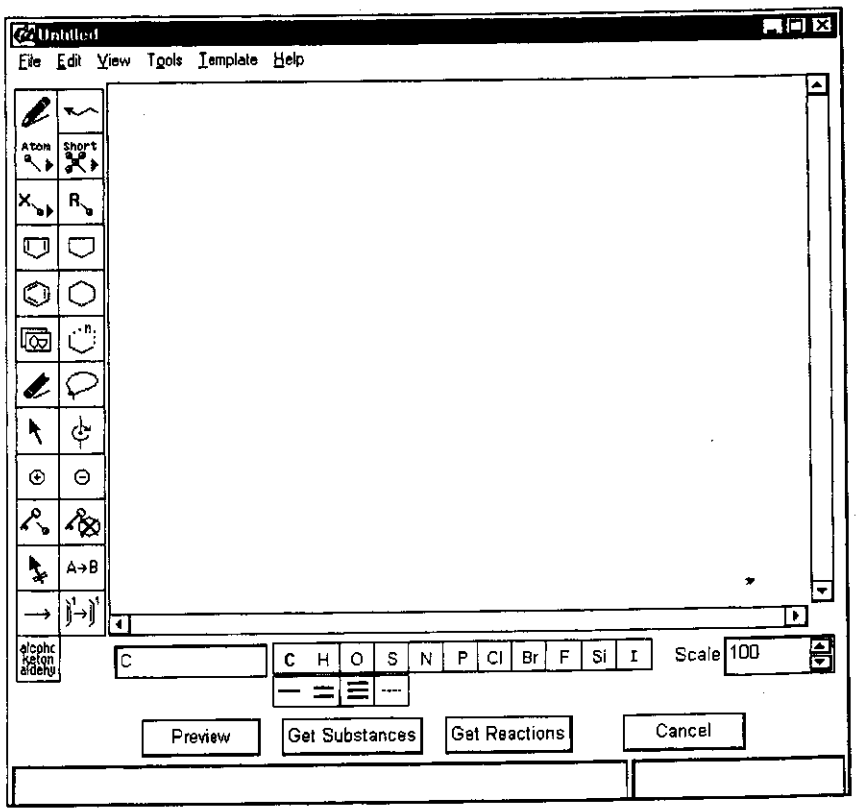



**Figure 3.2** Structure drawing screen. After the structure is drawn, the appropriate button at the bottom is clicked. **Get Substances** has two options (Figure 3.3). SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

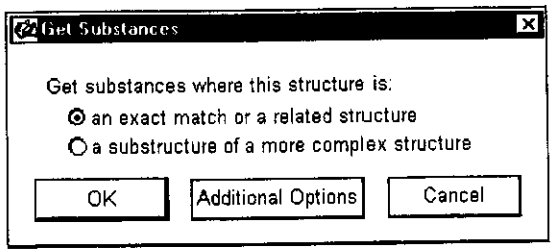
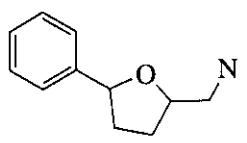
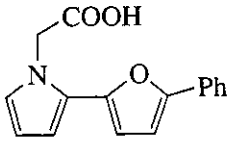



**Figure 3.3** Options within **Get Substances**. SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

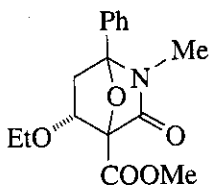
Many of the issues that may be confronted are best illustrated through an example. Thus, suppose information on the primary amine represented in part by structure (**2**) is required. While there are a number of issues to consider before the search is commenced (e.g. stereochemistry?, exact substance?, salts?, substructures?), nevertheless, suppose that the structure (**2**) is drawn and is searched for 'an exact match or a related structure'. When this is done, no answers are obtained. But what has been searched and what should the user do next? Remember, the application of scientific method to literature retrieval requires understanding of what has been done and then, rational design of the next experiment!



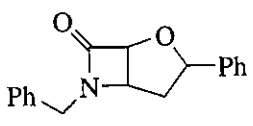

(**2**)



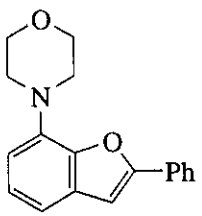

(**3**)



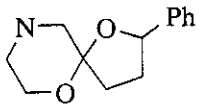

(**4**)




(**5**)




(**6**)




(**7**)

The user may then think: 'OK, the exact structure is not known, so perhaps it will be valuable to find substances related by structure', whereupon, 'a



substructure of a more complex structure' (Figure 3.3) is chosen. In this specific case, a message appears: 'Your structure is too general. SciFinder Scholar will not be able to complete your request to **Get Substances**'. SciFinder Scholar then suggests an 'autofix' option. What is that? What is the next step?

If the six-membered ring is locked (see below) and if this modified structure is searched, almost 400 answers are obtained and some representative structures are (**3**) to (**7**). Why are some of these retrieved and more importantly, what should now be done to obtain more specific answers? (e.g. the trimethoxyphenyl and various chlorophenyl derivatives of the primary amine (**2**) are some of the more immediately related analogs among the 400 answers.)

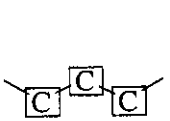
It is apparent that it helps if the searcher understands a number of issues in order to search for structures effectively, and answers to all the aforementioned questions are given in the following sections.

### *3.4.1 Drawing Structures*

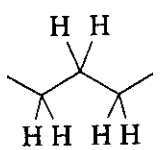
When first using the program, the user is advised to scroll through the pull-down menus at the top and the structure drawing tools at the vertical bar on the left of the screen (Figure 3.2). Most of the options are self-explanatory although a few require further comments.

First, clicking the button **Short** opens up a screen, in which chemists will recognize the most commonly encountered functional groups. When shortcuts are chosen, only those precise groups will be searched and answers with further substituents on these groups will not be retrieved even in a substructure search.

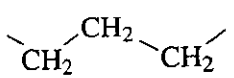
Second, the **Lock Out Tool** is used to block further substitution in a substructure search at the position locked. For example, if a five-carbon chain is drawn and the locking tool is clicked on the three middle carbons, then structure (**8**) results. Further, substitution at these carbons is thus prevented and the search result is the same as if structures (**9**) or (**10**) had been drawn.



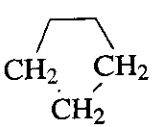

(**8**)




(**9**)




(**10**)




(**11**)

Third, the **Ring/Chain Lock Out Tool** is used to block formation of rings. Thus, *when rings are drawn in SciFinder Scholar, the atoms default*

*to isolated or embedded values*, and *when chains are drawn, the chain atoms and bonds default to ring or chain values*. Accordingly, if defaults are not overridden, a substructure search on structure (**8**) would include cyclopentanes (**11**) as answers.

Presently, it is not possible to prevent formation of rings just at one position. Thus, clicking the tool on any atom in a ring isolates the entire ring and clicking the tool on any atom in a chain isolates the entire chain. However, separate rings and different chains may be locked independently. It should however be remembered that specific types of answers may either be selected manually or may be extracted through any number of structure analyze or refine tools (Section 3.5) that may be implemented once an initial answer set has been obtained.

The locking tools do not impact on searches for the exact or related structures (Section 3.4.2) because in these cases, hydrogen atoms, or hydrogen isotopes or charges, are inserted automatically. So, if structure (**12**) (Figure 3.4) with the rings or chains not locked is drawn and if an 'exact match or related structure' search is performed, then answers of the type (**13**) and (**14**) will be retrieved. If a substructure search is performed, answers of the type (**15**) and (**16**) will also be obtained and it should be noted that retrieval of structure (**16**) occurs only because the structure search defaults for rings (isolated/embedded) and for chains (ring/chain) have been applied. However, if the rings/chains in the structure drawn are locked, then answers

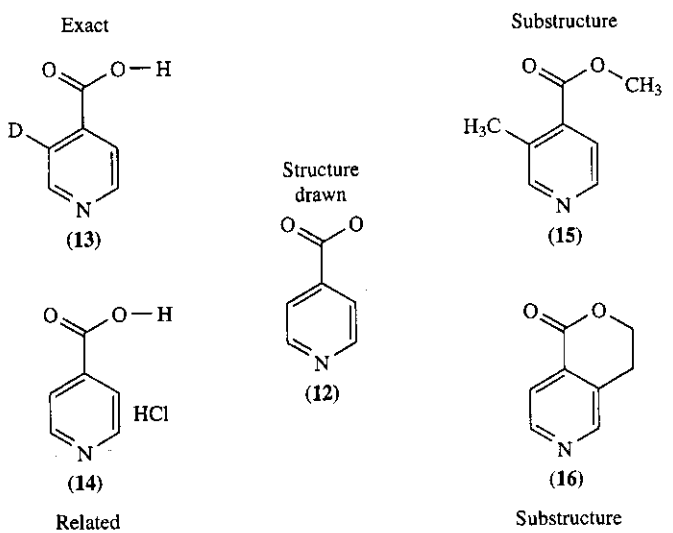



**Figure 3.4** Types of answers retrieved from structure searches.



of the type (**16**) will not be retrieved in a substructure search. It helps if users think carefully about how locking rings/chains will affect the search, and indeed, use of this tool should be *minimal in the initial search* since it may be better to implement subsequent tools to revise answers. However, the locking tool may be needed to overcome the issues of system limits (see following text) and in these cases, there may be times when more than one query may be needed to allow for all required options.

Finally, the icons at the bottom of the vertical palette (Figure 3.2) are used in drawing queries for chemical reactions (Chapter 6).

Structure drawing in SciFinder Scholar is very intuitive and structures may be drawn with little instruction. Users should also keep in mind the ability to copy a CAS Registry Number from *any record* within Scholar and to paste it into the structure drawing screen. To do this, the CAS Registry Number is carefully highlighted with the mouse, then copied (e.g. through the pull-down **Edit** menu or **CtrlC**). The structure drawing screen is then obtained, and when **Paste** is employed, the structure appears on the screen. Depending on the substance being copied, many of the hydrogen atoms in the original substance may be removed in this **Copy** and **Paste** operation. It is a simple matter to note carefully what has been copied and to make the required changes through the usual structure drawing tools.

### *3.4.2 Searching Structures*

Structure queries may be searched either for 'an exact match or a related structure' or for 'a substructure of a more complex structure'. The choice made depends on the intention of the search, and some of the issues have been mentioned in reference to structures (**2**) to (**7**). What is important is that the scientist understands the differences and applies the options that not only will best solve the immediate problem but also will lead to new discoveries.

#### Exact Match or a Related Structure

In drawing the structure query, it is not necessary to assign hydrogen atoms, since the exact or related structure search process automatically inserts hydrogen atoms at vacant positions. Further, while variable bonds (e.g. unspecified bonds) are allowed, variable atoms (e.g. generic groups Q, A, X, M, or R groups) are not allowed in the query.

Answers include all stereoisomers and isotopic substances and multicomponent substances where the exact structure is one of the components. Examination of entries in Appendix 5 will suggest to users when multicomponent substances are important in the search. For example, many biologically

active compounds are either organic acids or bases that are insoluble in biological fluids, and one of the key considerations in medicinal chemistry is to develop biologically compatible derivatives. This may be achieved either through formation of salts or through complexation with substances like cyclodextrins. Since salts and complexes are indexed as multicomponent substances with the biologically active substance as one component, it is apparent that searching for 'a related structure' will readily retrieve an answer set containing biologically compatible derivatives.

Searches for 'an exact match or a related structure' nearly always proceed to completion, and the user does not need to be concerned with any of the issues involved with substructure searches (Section 3.4.2). However, if the query contains a structure that is present as a component in many multicomponent substances (e.g. styrene is a monomer in almost 60 000 polymers), then large answer sets may be retrieved. In these cases, before the search is conducted, some filters[2] may be activated through the pull-down menu options (**Preferences** then the **Explore** tab). Alternatively, the button **Additional Options** (Figure 3.3) may be clicked and the same filters may be chosen.

### A Substructure of a More Complex Structure

Many additional issues arise when a substructure search is performed and now it helps to understand a few of the basics of the structure search process.

**The Screening Issue.** SciFinder Scholar performs structure searches in steps. First, substances are screened; second, potential answers are checked through atom-to-atom matches with the query; third, acceptable answers are sorted (grouped into certain classes); and finally, the answers are displayed.

The screening process mainly identifies structure fragments, some atom connectivities, and bond and ring types in the query, and matches these features with substances in the database. If the query is very general, then too many *potential answers* pass the screening process and the query needs to be modified. When this occurs, SciFinder Scholar gives a warning message: Your structure is too general. SciFinder Scholar will not be able to complete your request to '**Get Substances**', and suggests an 'autofix' option. If chosen, the locking tool is applied to all rings and chains, which means that further ring fusion is not allowed and that chains are specified as chain values only. If this process would remove the wanted answers, careful consideration needs to be given to drawing a different structure or to locking manually only parts of the structure.

[2] The options are to include/exclude multicomponent substances, or to specifically include/exclude incompletely defined substances, isotopes, metal-containing substances, mixtures, or polymers.



Note that the warning message relates to the screen and search stages and *not* to the answer stage. For example, a substructure search on structure (**2**) may exceed the system search limits, even though there may be very few answers. The point is that this structure contains mainly carbon–carbon bonds in quite common rings, and many substances will pass the screen process through which only a few structure fragments, which in any case occur very commonly, would have been identified.

If a structure query is 'too general', then more precise queries are obtained by locking rings or chains, by being more specific with atoms in generic groups (if used), or by adding further atoms to the query (Table 3.2). If the latter cannot be done so that all wanted structures are still retrieved, then a further option is to build and search separate structures. It is to be noted that while adding hydrogen atoms to the query blocks the substitution and hence reduces the number of answers in substructure searching, adding hydrogen atoms to the query may not affect the screening process and so may not help

**Table 3.2** Options to broaden or narrow substructure searches

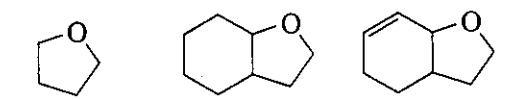
| Narrowing answers in substructure searches | Broadening answers in substructure searches |
|---|---|
| Specify more atoms in structure, i.e. build a more complex structure | Specify fewer atoms in structure; leave as many positions open as possible |
| Isolate ring(s) | Allow isolated/embedded rings (the default) |
| Restrict atoms and bonds in chains to chain-only values | Allow chain atoms and bonds to have ring or chains values (the default) |
| Block substitution, for example, by addition of hydrogen in vacant positions | Remove hydrogens |
| Define atoms more precisely (e.g. restrict generic groups) | Use generic groups A, M, Q, X or allow atom variables (R groups) rather than use specific atoms |
| Define bonds more precisely | Allow unspecified bonds |
| Change **Preferences** to exclude some **Explore** options, e.g. exclude multicomponent substances | Include further options under **Preferences, Explore** |
| Use options under **Analyze Substances** as a guide to obtain more precise answers | |
| Use **Refine Substances** and thus retrieve a subset of the answers on the basis of part structure drawn | |
| Use **Refine Substances** and limit answers by **Property Data** | |

*Note*: A large *substance* answer set may be acceptable and *references* may be reduced either through one of the CAS Role filters or through **Analyze References** or **Refine References**.

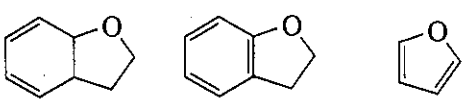
to overcome system search limits. Nevertheless, users are encouraged to add hydrogen at positions where the search definitely requires the substitution to be blocked.

**The Resonance Issue.** An additional issue relates to what constitutes an 'answer' and a principal consideration is the application of the concept of resonance in valence bond theory (Appendix 6). For example, while structure (**3**) contains the substructure (**2**), the user may have wanted answers only where the ring in structure (**2**) is saturated (i.e. a *tetrahydro* furan derivative). SciFinder Scholar interprets queries in the more general way; so, if a substructure search on structure (**2**) is requested, answers will include structures of the type (**3**) and (**6**) because both contain the atom connectivities in the query. The additional question here is the nature of the bonding, and it mainly depends on whether the locking tool has been used as to how bonds are interpreted in the search.

So, if a substructure search is requested on a 'tetrahydrofuran' (**17**) and if the ring is not locked, then SciFinder Scholar needs to allow for structures (**18**), (**19**), and (**20**). Consistent with the policy to provide more comprehensive answer sets and to allow the user to make evaluations in potentially marginal cases, SciFinder Scholar also allows for structure (**21**). However, since bonds in resonance structures are defined differently from single or double bonds (Appendix 6), SciFinder Scholar has to make a generic definition for the bond in the query (**17**) to allow for all possibilities. As a second six-membered ring may be fused onto (**17**) (a dibenzofuran structure), and then furan structures (**22**) need to be considered, and structures (**3**) and (**6**) are presented as answers. If the ring is locked, then SciFinder Scholar does not have to allow for (**21**), and mainly tetrahydrofuran structures result.




(**17**) (**18**) (**19**)

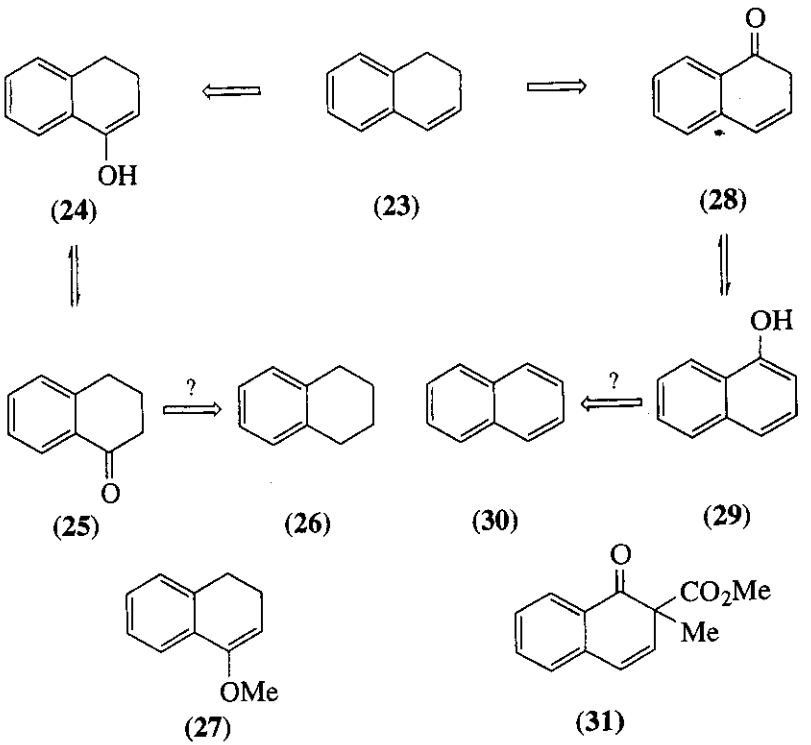



(**20**) (**21**) (**22**)

**The Tautomerism Issue.** Complications also arise with tautomers (Appendix 6). For example, if a substructure search is requested for structure (**23**), then



a possible answer is (**24**), which is a tautomer of (**25**). It could then be argued that (**25**) contains the substructure (**23**), but then how is (**25**) related to (**26**)? There clearly is a problem with answers of the type (**26**) to a query based on (**23**), and a first thought may be to exclude the complication created by the tautomers (**24**)/(**25**). Such an exclusion leads to the question of how to treat (**27**), which clearly has the substructure (**23**) and is related to (**24**)!




In a similar way, structure (**28**) contains the substructure (**23**), but the tautomer issue arises with (**29**), which leads to a query concerning (**30**). Here, it may be argued on chemical grounds that (**28**) will mostly be in the form (**29**), but how about structure (**31**)?

In summary, the substructure search algorithm within SciFinder Scholar needs to consider substances based on (**26**) and (**30**) when interpreting a query (**23**), and there arise possibilities perhaps beyond the initial expectations of the user. Consistent with its policy of providing alternatives and allowing the user to choose, SciFinder Scholar provides a comprehensive answer set, which in this case may include naphthalene and tetrahydronaphthalene derivatives (i.e. substances of the types (**30**) and (**26**), respectively). While

the user may eliminate unwanted answers manually or by modification of the structure query (e.g. by adding hydrogen or other atoms at key positions), SciFinder Scholar, in fact, offers very powerful structure analyze and refine options (Section 3.5) that solve the problem much more elegantly!

### Performing the Search: Preview or Get Substances

Once the initial structure has been built, the user chooses between two buttons, **Preview** and **Get Substances**. If system-defined variables and user-defined variables (e.g. X and $R_1$, respectively, in Figure 3.5) are present in the structure, all the options under **Preview** will be active (Figure 3.6).

**Preview a sample of: Answers** performs a search of a sample of the database, produces some *actual answers*, and projects the result of a full

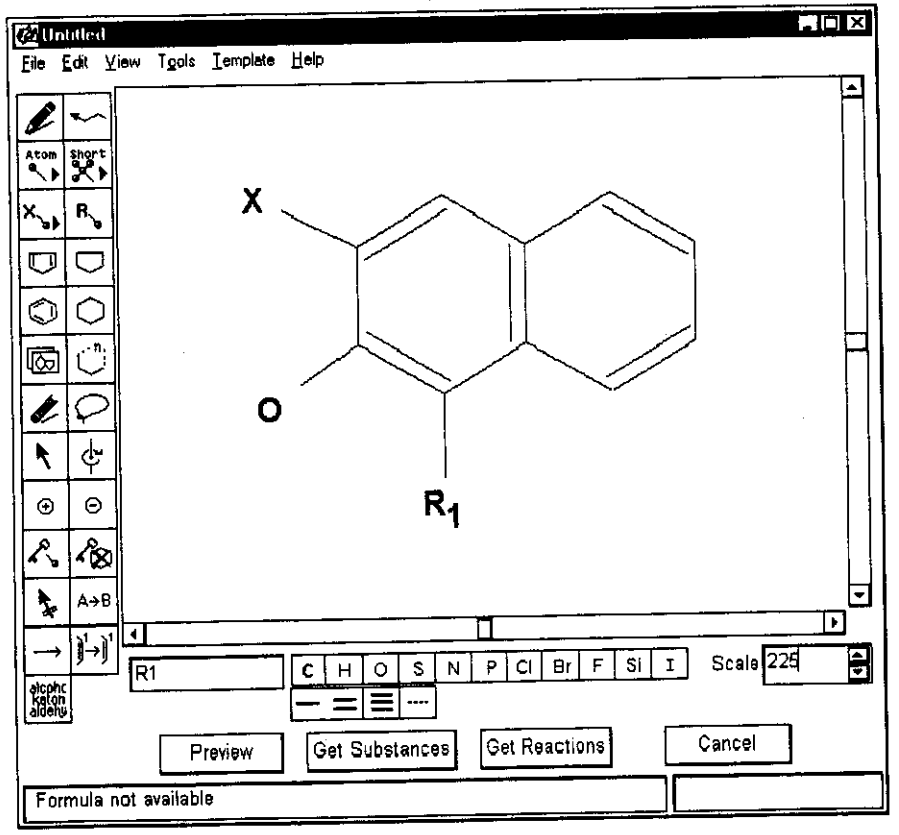



**Figure 3.5** Structure query ($R_1$ = O, C; X = any halogen atom). SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.



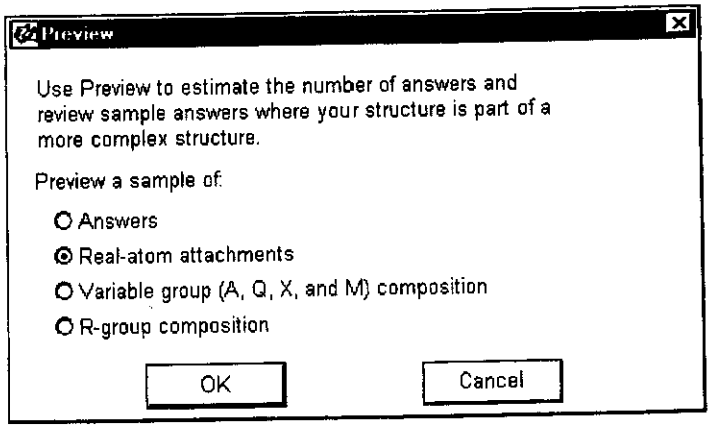



**Figure 3.6** Options under **Preview**. SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

search. In the example (Figure 3.5), 10 sample answers are displayed and the system projects an answer set of around 450 substances. **Preview a sample of: Real-atom attachments** gives another screen and when any node is clicked, the various atoms attached to this position are presented (Figure 3.7). The screen also indicates approximately how many answers will be retrieved in the full search. In a similar manner, **Preview a sample of: Variable Group** and **Preview a sample of: R-group composition** indicate the percentages of the different types of variable groups. Once again, SciFinder Scholar is guiding the searcher through indicating potential outcomes from which choices may be made. However, *provided the answers are of the general type required, it is not essential for choices to be made at this stage*, as these and other analysis tools may be performed later on in the full answer set. **Preview** is particularly important for SciFinder users who have 'task packages' as it enables them to assess possibilities at no cost.

When **Get Substances** (Figure 3.5) is clicked, the two choices are illustrated in Figure 3.3. As noted in Section 3.4.2, 'an exact match or a related structure' affords answers in which all positions are blocked with respect to all atoms other than hydrogen and its isotopes.

If 'a substructure of a more complex structure' is chosen, then a full substructure search is undertaken and the answer set is obtained (Figure 3.8). In this case, more than 600 answers are found and the user has a number of options. Answers may be printed or saved (e.g. through the icons at the top of the screen) and individual records may be checked (check box) or displayed in full (microscope). Many substances also have links to 3D Models or

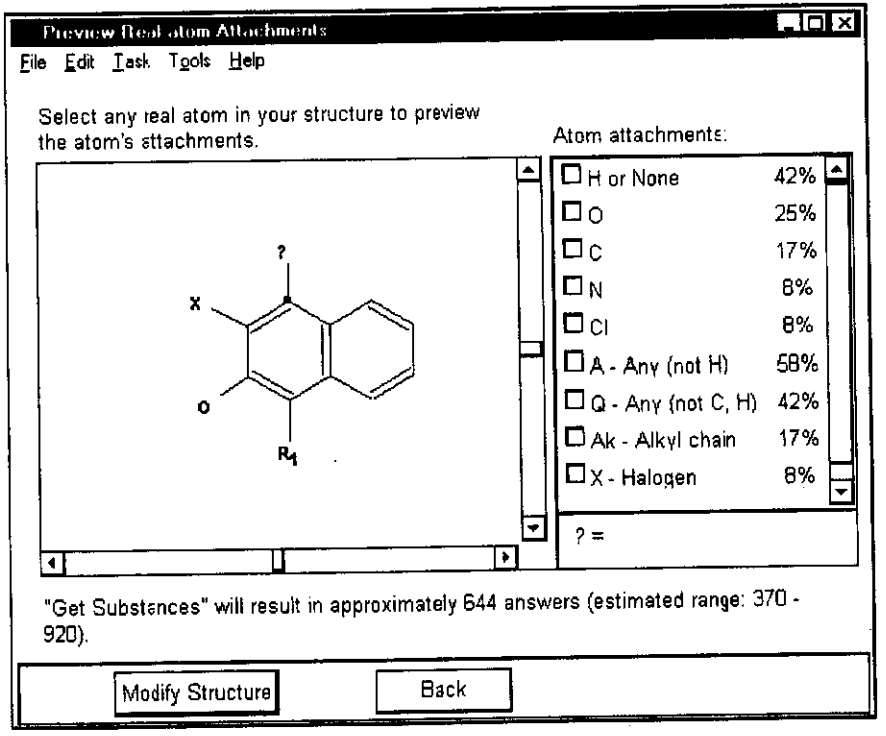



**Figure 3.7** Display of substituent atoms on substructure query through **Preview: Real-atom Attachments**. (See also Figure 3.10.) SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

to CHEMCATS or CHEMLIST. However, the '**Get References**' and the '**Analyze or Refine Substances**' buttons provide the key options, and the operations are discussed below.

## 3.5 Analyzing and Refining Substance Answer Sets

Before references are obtained, it is usual to review the list of substances and one option is to do this manually. However, if many hundreds of structures match the structure query, this manual process may be tedious. Further, inexperienced users may at first not fully understand some of the answers and why they have been retrieved. It is just like a scientific experiment in that at first the user may simply not understand what the experiment has produced! The solutions come through an understanding of the content of the file and of the opportunities that SciFinder Scholar provides to evaluate the data.



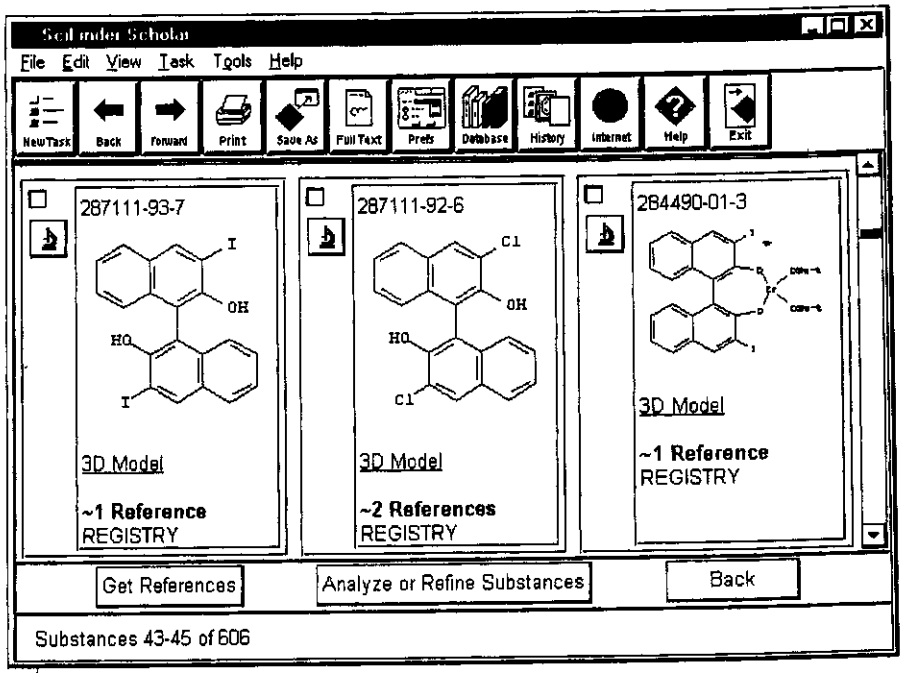



**Figure 3.8** Some answers from substructure search of query Figure 3.5. (Note the options: Check box, Microscope, **Get References, Analyze or Refine Substances.**) Copyright the American Chemical Society and reproduced with permission.

SciFinder Scholar offers analyze and refine options for substances in a manner similar to the way it offers these options for references (Chapter 2, Section 2.5.4). That is, the option to analyze substances gives histograms that group answers according to certain specifications, while the option for refine allows precise inputs to be made directly. As with analyze/refine for references, the experienced user usually first takes advantage of **Analyze Substances**, since this gives an idea of the types of alternatives that may be followed before actually making a decision. Indeed, often it is advisable for the searcher to set up more general queries and then to be guided to more specific structures through analyze or to explore different options through refine. Thus, although the group $R_1$ has been defined as either oxygen or carbon in Figure 3.5, it may have been better to leave the position vacant and then to analyze attachments at this position after the initial answers are obtained. Why limit answers, since perhaps, a more general search followed by analyze may have turned up some additional and very interesting substances?

### 3.5.1 *Analyze for Atoms*

Of the five analyze options (Figure 3.9), the first three provide information on atoms. **Real-atom attachments** (Figure 3.10) displays a list of the actual atoms attached at the various open positions in the query structure and the frequency with which they occur. **Variable group Composition** (Figure 3.11) displays the frequency of actual atoms in the structures at positions where variable groups A (any atom other than hydrogen), Q (any atom other than hydrogen or carbon), X (any halogen), and M (any metal) have been used in the structure query. **R-group Composition** (Figure 3.12) displays similar options where a user-defined variable (R group) is in the structure query. When this is done, the outcome is shown in a display sorted by frequency. One or more of the boxes may be checked, in which case the user proceeds to **Get Substances** and thus obtains a refined answer set.

The intention of these options is to provide information on the types of substituents present in the answers. Knowing the substituent atoms, and their frequency, greatly helps the user decide on the next step since the nature of the problem is understood before the decisions need to be made.

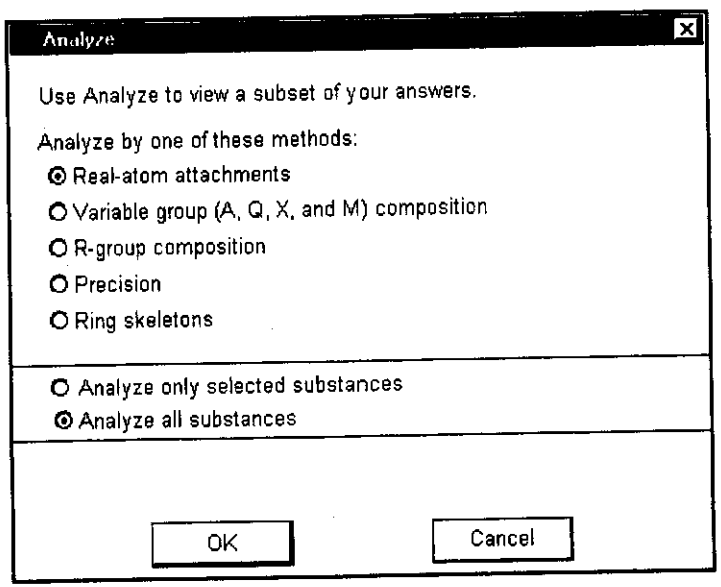



**Figure 3.9** The five options under **Analyze Substances**. The first three relate to atoms, **Precision** relates mainly to bonding definitions, and **Ring skeletons** identifies ring systems. SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.



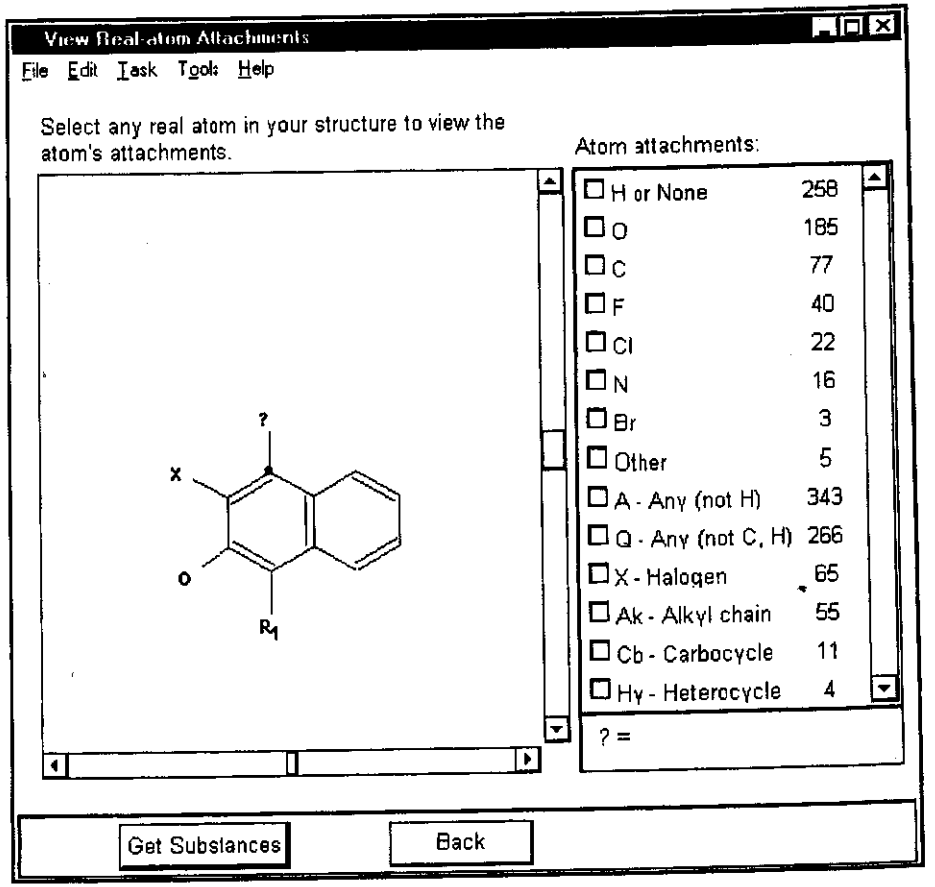



**Figure 3.10** **Real-atom Attachments** at 4-position for answers from substructure search, Figure 3.5. Similar tables for other free positions may be obtained by clicking on the appropriate atom. This display differs from that in Figure 3.7 since the analysis has been done on *actual answers* obtained. Figure 3.7 is a preview and gives projections of outcomes only. SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

Sometimes, research areas are prompted by screens like these. For example, analysis by **R-group Composition** (Figure 3.12) indicates that only three of the substances have heterocyclic groups, and the scientist's knowledge of heterocyclic chemistry and interest in these naphthols may spark further interest in these substances. Research in academic laboratories is very much driven by curiosity, and the opportunities to develop new research programs on the basis of knowledge of prior literature should never be overlooked.

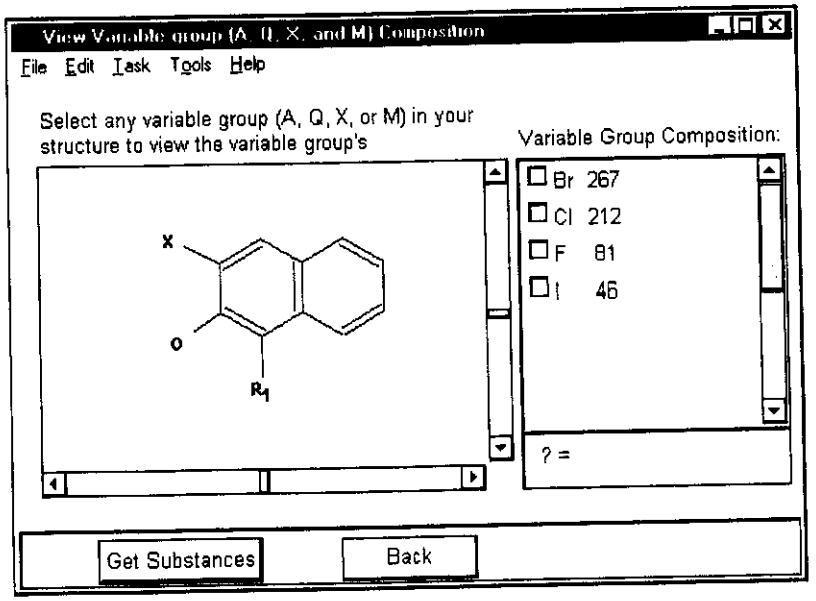



**Figure 3.11 Variable group Composition** (X) for answers from substructure search, Figure 3.5. SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

If, indeed, 2-naphthols with heterocycles at the 1-position are relatively unknown, then is this because others have explored them and discarded them as uninteresting or because others have never investigated what may turn out to be a new series of compounds with very interesting properties? On the other hand, research in industrial laboratories is in part driven by novelty, and analysis tools of these types may quickly alert scientists to new areas. Knowing what is not in the literature may at times be as significant as knowing what has been reported. SciFinder Scholar should be considered not only as a source of information but also as a creative research tool!

### *3.5.2 Analyze by Precision*

The other two analyze options have quite different functions. The first, **Analyze** by **Precision**, does exactly that – it looks to see how precisely the various answers match the initial query. This may seem an irrelevant issue, since after all, are not structures precise entities and hence should not all answers precisely match the query? In fact, one example of the problem of what precisely constitutes an answer was discussed earlier in relation to the complex chemical concept of tautomerism (structures **23**–**31**).



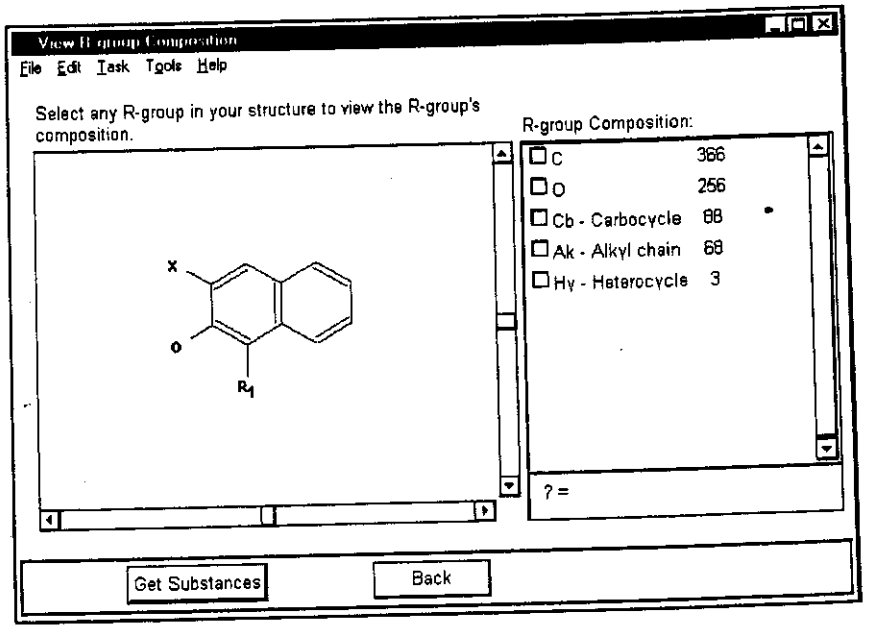
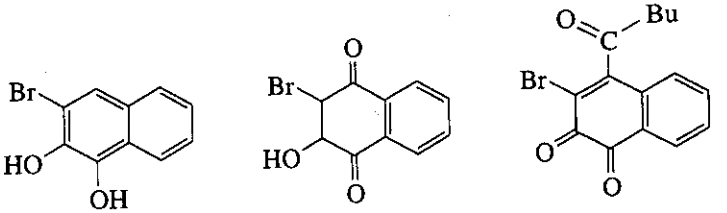



**Figure 3.12 R-group Composition** ($R_1$) for answers from substructure search, Figure 3.5. SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

Thus, a substructure search on the structure query (Figure 3.5) gives more than 600 answers including structures (**32**) to (**34**). Structures (**33**) and (**34**) may have been unexpected but they arise because of the tautomerism issue. The solution is to **Analyze** by **Precision**, and the results of the precision analysis are given in Figure 3.13. What constitutes 'closely associated tautomers' or 'loosely associated tautomers' depends very much on the question being asked, and the options merely need to be viewed to assess their relevancy. In this case, structures (**33**) and (**34**) appear as 'closely associated' and 'loosely associated' tautomers, respectively.




(**32**) (**33**) (**34**)

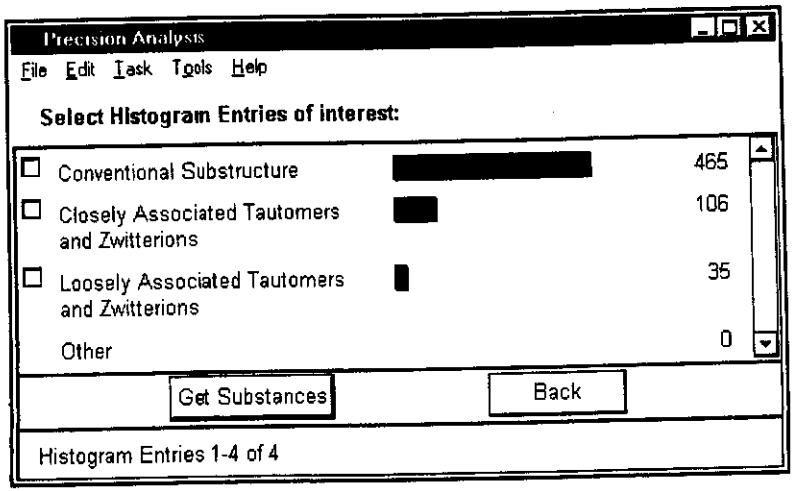


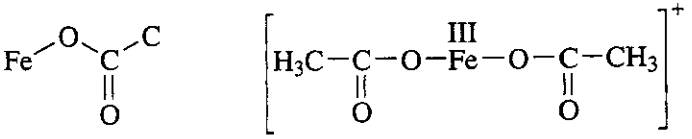

**Figure 3.13** Histogram for **Analyze** by **Precision** from answer set in Figure 3.8. SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

The other main case, where a key question is what constitutes an answer, is best illustrated by an example. That is, if the structure query (**35**) is drawn, then would substances (**36**), (**37**), and (**38**) be reasonable answers from a substructure search? The point is that ferric acetate (**37**) is indexed as a salt, whereas the other compounds are treated as coordination compounds (Appendix 5.3); yet, in a sense, all contain the part structure (**35**). In the same way that organic chemists need to deal with tautomerism and resonance, coordination chemists need to deal with what constitutes a salt and what constitutes a coordination compound. So, when a substructure search on query (**35**) is performed, SciFinder Scholar needs to allow for all possibilities and usually does so by effectively *ignoring the bonds to the metal atoms*. But this is not always the case as in some cases, for example, metal-containing porphyrins, the algorithm does not ignore the bond to the metal. However, in being comprehensive, issues of precision again arise.

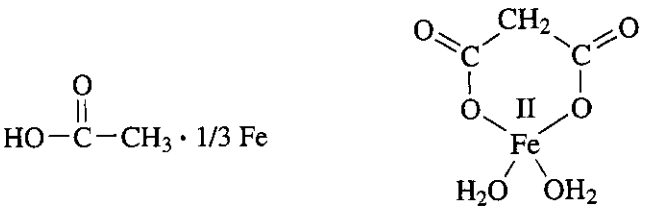
**Analyze** by **Precision** is used to address the issue associated with the way SciFinder Scholar interprets salts/coordination compounds. For example, because of the fact that the bonds to the zinc atom in (**39**)[3] are initially ignored, a substructure search produces more than 2000 answers. All answers have a carboxylic acid group and a zinc atom (but not necessarily directly

[3] The intention was to find Zn-biscarboxylate chelates. When users are unsure of how the bonding may be represented in the database, it may be advisable to unspecify the bonds (as shown in (**39**)).



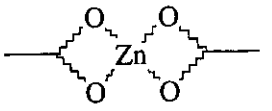


(**35**) (**36**)



(**37**) (**38**)

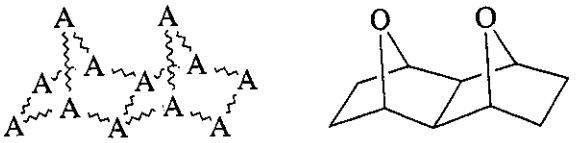
bonded), and many are multicomponent substances. However, **Analyze** by **Precision** lists 10 substances under 'Conventional Substructure' and these are exactly what is wanted (i.e. are zinc biscarboxylates of the type (**39**)).



(**39**)

### *3.5.3 Analyze by Ring Skeletons*

**Analyze** by **Ring skeletons** provides three options (Figure 3.14). They become more specific from **Ring skeleton only** through **Ring skeleton with atoms and bonds**, and it may be advisable to take them in stages. Note that **Analyze** by **Real-atom Attachments** (Figure 3.10) looks simply at atoms attached to the initial query irrespective of whether they are in rings or chains, while **Analyze** by **Ring skeleton** more specifically looks at ring structures and thus provides important additional information.



(**40**) (**41**)

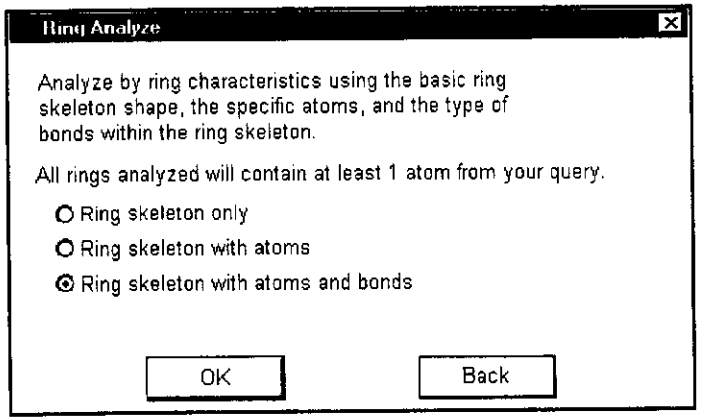



**Figure 3.14** Choices for **Analyze** by **Ring skeletons**. SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

For example, if substances of the ring system (**40**) are of interest, a substructure search would first be attempted. If the rings are not locked, then understandably, the query is too general, but when the rings are locked, more than 3300 substances are retrieved. **Analyze** by **Ring skeleton with atoms** gives more than 30 systems (Figure 3.15), and research opportunities may now be considered. Of course, these 30 ring systems are restricted in that ring fusion is blocked in the initial query; so, if higher ring systems are required, the searcher may then start with a more specific query, for example, ring system (**41**). A substructure on this query completes without locking the ring and gives more than 600 structures, which on further analysis, contain just over 130 ring skeletons. Some of these contain fused porphyrins or crown ethers, and chemists would understand that these substances are of interest because the rigid ring system (**40**) provides an excellent scaffold for the attachment of functionally important moieties.

These analyze options are important tools that are available to the creative searcher. Retrieval of information is a real intellectual challenge!

### *3.5.4 Refine Substances*

The final option in Figure 3.8 is **Refine Substances** and when chosen, the options of refinement by **Commercial Availability, Chemical Structure**, or **Property Data** (Chapter 4, Section 4.6) are presented. The first retrieves only those substances that are listed in CHEMCATS (Chapter 1, Section 1.9), while the second places the user back into the structure drawing screen



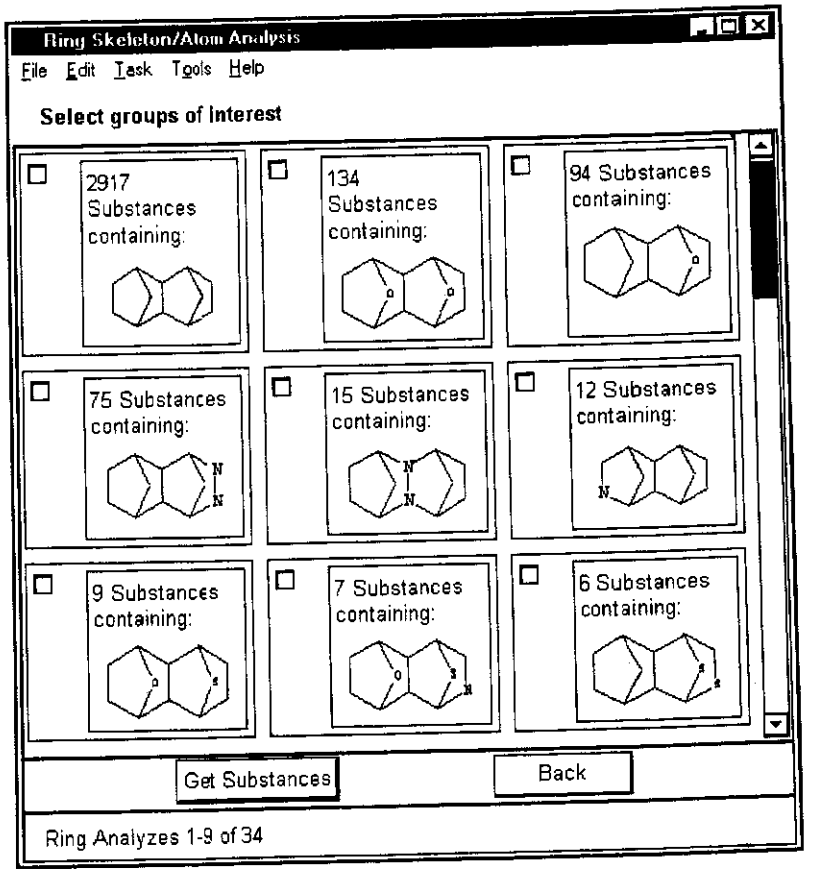



**Figure 3.15** Result of **Analyze** by **Ring skeletons with atoms** for substructure query on structure (**40**). SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

(Figure 3.2). Any new structure may be drawn (e.g. to check for metal-containing substances, the structure refinement query would just contain the node, M!) and when **Get Substances** is entered, SciFinder Scholar effectively searches this new structure within the original answer set.

This subset of the original answers may be further analyzed or refined, may be printed or saved, or references may be obtained. At any stage, the user may return to the original answers and execute further operations.

Note that refine works at the 'and' level, that is, produces answers that have the original requirements 'and' the refine requirement. At present, it is not possible to 'not' out structures, that is, to produce answers from the initial set

that do not have certain structure components. Removal of unwanted answers has to be done manually, that is, by clicking only those boxes next to the structures that *are* required.

When the final answer set is obtained, the next step is to **Get References**. The outcomes are discussed in Section 3.9.

## 3.6 Explore by Substance Identifier

When **Explore by Chemical Substance**, then **Substance Identifier** is chosen, the screen (Figure 3.16) appears, and names or CAS Registry Numbers of up to five substances may be entered. The search algorithm first looks for an exact match between the name entered and a complete name in the database. If an exact match is achieved, the answer is displayed. If, however, an exact match is not achieved, the algorithm attempts to give some reasonable answers by looking at the parts of names.

Unless the user is very proficient in nomenclature, name searches generally succeed best when common or trade names are known. However, even here,

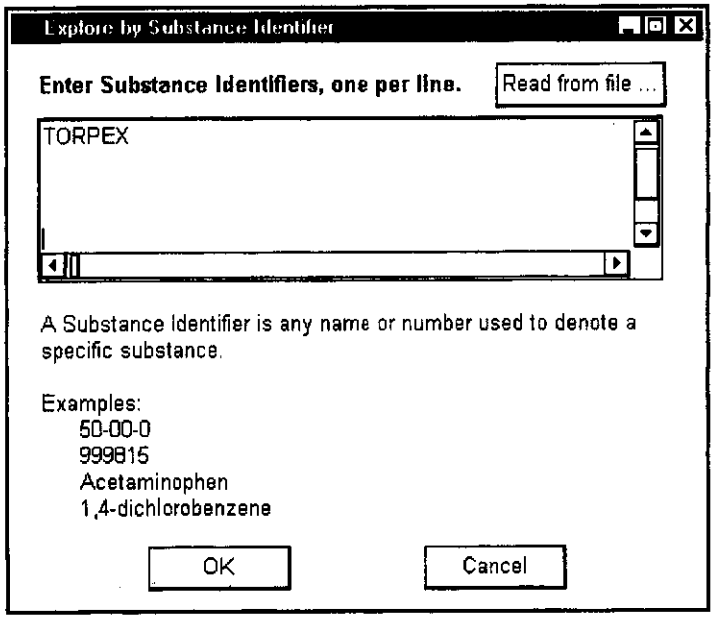



**Figure 3.16** Screen for entry of substance names or CAS Registry Numbers. SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.



there may be issues to address. For example, there are around 20 records for 'calcium sulfate' and its hydrates, and a number of these substances may be of interest. It is important that the searcher tries a few alternatives and if the required substance is not obtained in the first search, some of the answers may suggest how closely related substances may be named.

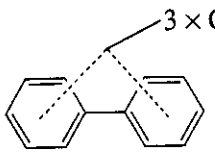
The point is that matching names in the substance database is a challenge, and users familiar with the complexities of searching for name fragments (e.g. those familiar with searching in the Basic Index and Chemical Name Segment fields in the REGISTRY file on STN) will appreciate the difficulty in writing an algorithm to search for all name possibilities. In summary, SciFinder Scholar performs very well when exact matches are achieved, but in other cases, the user needs to explore the alternatives. Often, it is a matter of exploring a few name-based possibilities, and if they fail or if answers do not retrieve all required derivatives (salts, hydrates, etc.), then alternative searches need to be explored.

An alternative way to find CAS Registry Numbers is to enter terms under **Explore by Research Topic**, and then to look through the answers.[4] For example, a search through **Substance Identifier** for 'acetone oxime' retrieves around 100 potential substances (search on acetone oxime does not give an exact match), and while these may be examined to find the exact substance, another way to find the CAS Registry Number is to enter 'I am interested in ''acetone oxime''' under **Explore by Research Topic**, and then to look through a couple of the records under the candidate 'as entered'. (However, an exact structure search should also be considered!) The CAS Registry Number, 127-06-0, readily becomes apparent. This raises the issue as to whether information on substances should be found directly through **Explore by Research Topic** or **Explore by Chemical Substance: Substance Identifier**. The many issues involved in answering this question are discussed in Chapter 5, Section 5.2.

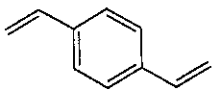
## 3.7 Explore by Molecular Formula

Many chemists are familiar with the process of finding a substance in print CA or CA on CD. Substances are first located by formula, then the examination of the list of possibilities by name eventually leads to the desired substance. Essentially, the same process may be followed in SciFinder Scholar, although the refinement is made through a structure rather than a name search.

[4] Many of the more common substances entered since 1985 have names used by the authors immediately following the CAS Registry Numbers.




(42)

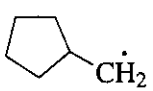


(43)

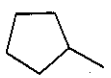
Second, if 'an exact match or related structure' search (Section 3.4.2) produces too many multicomponent substances and if only the exact substance is required, a formula search for the exact substance is first performed. If needed, the substances are then refined by a structure. For example, 'an exact match or related structure' search for *p*-divinylbenzene (**43**) gives almost 1000 substances, and the relatively large number is due to the fact that the substance is a component in many polymers. On the other hand, a search on molecular formula C10 H10, followed by **Refine Substances** with the structure (**43**) gives around 10 substances (i.e. the required substance, CAS Registry Number 105-06-6, and various isotopic forms).

Third, isotopes of hydrogen may be searched with deuterium (D) and tritium (T) in the formula (but isotopes of the other elements may be searched only through the element symbol). The required substance then needs to be selected from the list of the substances retrieved. Accordingly, a quick way to find the deuterated compound shown in Appendix 5.1.2 is through a molecular formula search: C10 H13 D N2.

Finally, formula searches may be the preferred starting option for the retrieval of reactive intermediates such as carbocations, carbanions, and radicals. For example, one way to find the cyclopentylmethyl radical (**44**) is first to search for the formula C6 H11, and then to refine substances with the structure query drawn as shown in (**45**).




(44)



(45)

It is helpful to understand all the search options to find substances and to try alternatives, particularly when initial strategies do not appear to retrieve the answers expected. Formula searches provide unique options, although it may be necessary to have a little understanding of how formulas are entered in the database, and Table 3.3 and Appendix 5 provide general rules and examples, respectively.



## 3.8 Special Cases

### *3.8.1 Multicomponent Substances*

Approximately one-tenth of the substances in REGISTRY are multicomponent substances, but they remain a source of confusion to many scientists mainly because the concept is rarely taught in college chemistry programs.

Multicomponent substances are more commonly encountered in mixtures, alloys, copolymers, and salts. The key factors are the way the components are listed in the molecular formula and structure fields, and many examples are given in Appendix 5.

Naturally, if a specific name for a multicomponent substance is known and is present in the database, the simplest option is to **Explore by Chemical Substance**, followed by **Substance Identifier**, and then to enter the name. For example, augmentin (Appendix 5.2.3) is easily retrieved through the trade name, although clearly there would be difficulties with searching for the systematic name!

Alternatively, if the formulas of the components are known, it is a simple matter to enter the formula and to see what transpires. For example, polypyrrole, the homopolymer of pyrrole, may be retrieved by name (e.g. polypyrrole) or by the formula search (e.g. (C4 H5)x).

However, if name or formula searches do not retrieve the required answers, then the power of structure searches through SciFinder Scholar can truly be illustrated. Thus, if it is known that 'torpex' (Appendix 5.2.3) is a mixture that contains aluminum, TNT, and trinitrohexahydrotriazine, it is a simple matter to draw the structures of these substances under the **Chemical Structure** option (Figure 3.18). If **Get Substances** and then 'an exact match or a related structure' is chosen, mixtures of just these three substances are retrieved. If **Get Substances** and 'a substructure of a more complex structure' is chosen, then six substances are retrieved including the components of torpex present in other explosive compositions.[5]

In general, multicomponent substances may readily be retrieved if the structures of the individual components are drawn. 'An exact match or a related structure' search may either produce answers that include just the components or may include these components in records in which further components are present. 'A substructure of a more complex structure' search produces answers in which the part structures are present and while they may be present in a single substance, multicomponent substances containing

[5] 'An exact match or a related structure' search on the structure of TNT is a more general option and subsequent analyze/refine tools may be used.

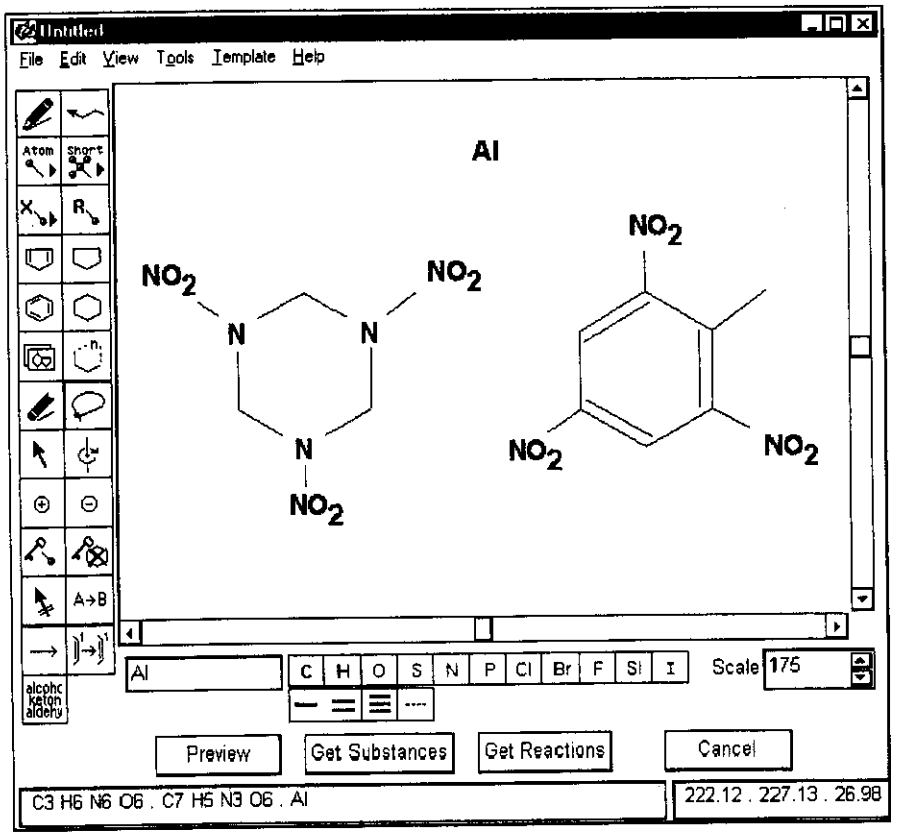



**Figure 3.18** Structure query for search on substances with the three components ('torpex'). SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

all the individual components will also be retrieved. If 'a substructure of a more complex structure' search produces unwanted substituents on the fragments or produces single substances that contain the fragments, then these may be excluded by drawing hydrogen atoms on the initial structure as required or may be eliminated from the answer sets manually or through analyze/refine options.

### *3.8.2 Structure Fragments*

There will be instances in which the strategy of retrieving multicomponent substances by drawing individual component structures on a single screen will not produce the required answers. For example, if the requirement is



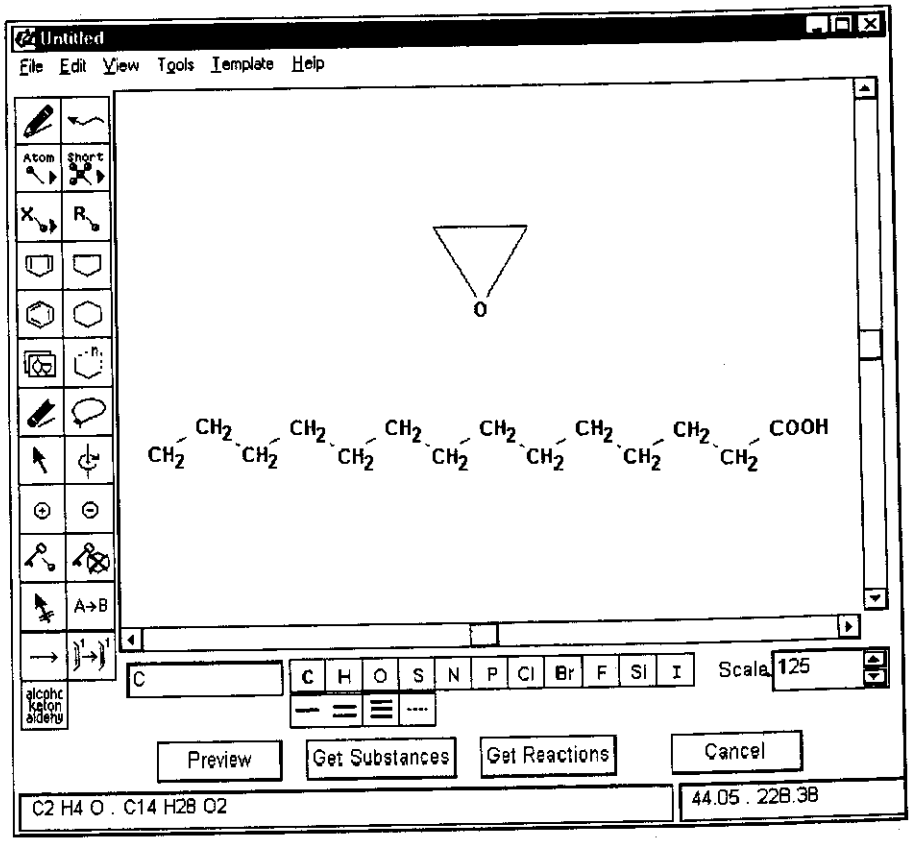
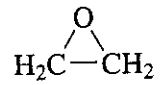
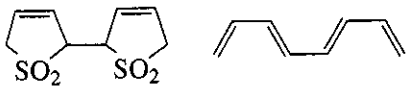



**Figure 3.19** Structure query to find multicomponent substances containing ethylene oxide and fatty acids. SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

to find copolymers in which one of the components is ethylene oxide **(46)** and a second is any straight-chain, saturated carboxylic acid with at least 14 carbons, then the query (Figure 3.19) will also retrieve answers in which the two part structures are in the same molecule. The way around this is to block substitution on the ethylene oxide (e.g. by attaching hydrogen), and with the ring locked, 'an exact match or a related structure' gives around 20 substances.



**(46)**




(47) (48)

Strategies of this type produce simple but elegant solutions to problems that cannot be solved by any other desktop information resource. The trick is to think creatively about the problem and not to be too focused, particularly in the early stages. For example, consider a request for substances of the substructure (**47**), which are required as precursors for the bis-diene (**48**). In turn, the focus of the study is to explore bis-Diels–Alder reactions (Scheme 1).

**Scheme 1**

One of the first questions to ask may be whether 1,3,5,7-tetraenes (**48**) really are desirable intermediates. May not the reaction occur through the two central double bonds? Indeed, is it necessary for the two rings in (**47**) to be directly connected? Would a single atom or several atoms between the two rings be acceptable, and if so, do all the atoms need to be carbons, or would other atoms be of interest? If the last question is answered in the affirmative, then how may the structure query be drawn?

When all these matters are considered, the user probably would simply **Get Substances** (substructure) from the query in Figure 3.20. That is, the most general search would be simply to find any substances with two sulfolene groups, no matter how they are connected. Such a search produces around 60 substances – a small enough answer set to look through manually, although analyze/refine tools may be used.[6]

## 3.9 Getting References

While at times, the substances alone are of interest, generally, the *references* are required. In **Explore by Chemical Substance**, the answers retrieved are chemical substances and when **Get References** is chosen, only the CAS Registry Numbers for the substances are used as search terms. The references

[6] All examples in this book are taken from actual queries raised with the author. In this case, analysis of some of the answers prompted a whole new research area, and it is a good example of how SciFinder Scholar may be used as a creative research tool.



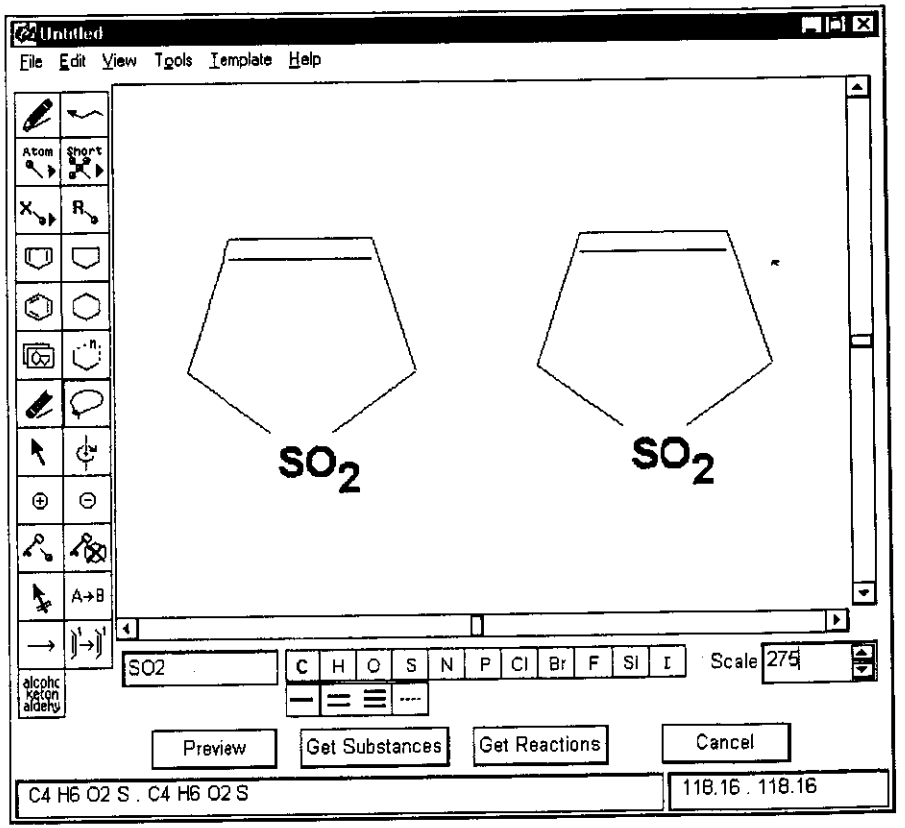



**Figure 3.20** Structure fragment query. SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

obtained are from CAPLUS and MEDLINE, and all the refine/analyze options discussed in Chapter 2, Section 5.4 or the **Get Related** options (Chapter 4, Section 4.5) may then be used in the normal way.

All of these refinements work at the Boolean 'and' level, so it is not possible to refine in a way that the CAS Registry Number is 'closely associated' with the topic. The exception to this is that when **Get References** is first selected, an option comes up to refine by a number of parameters (Figure 3.21). If chosen, these refinements effectively search the CAS Registry Number(s) and the terms associated with the refinement in CAPLUS at the 'closely associated' level. Since CAS Registry Numbers are in a separate field in MEDLINE (see Figure 1.5), a different algorithm is used in the crossover to MEDLINE.

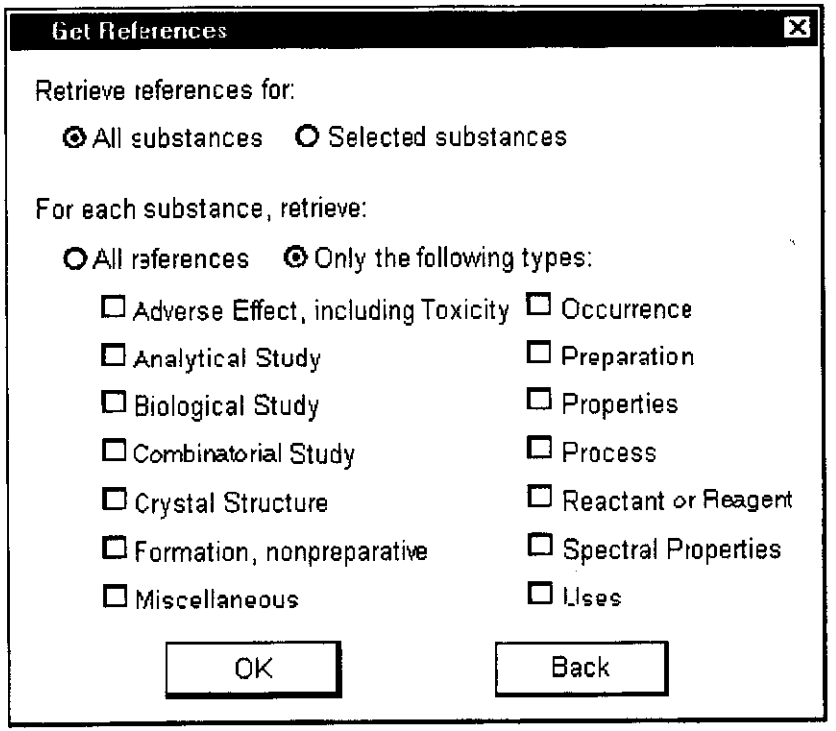



**Figure 3.21** Options for retrieval of specific types of information on substances. SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

It is also necessary to understand that policies for the indexing of substances are different in CAPLUS and MEDLINE (e.g. see Figures 1.3 and 1.5). In general, CAPLUS contains indexing at the most precise level (individual substances), whereas MEDLINE may index specific substances under the more general (or 'parent') substance. This is one reason that there are relatively few CAS Registry Numbers (almost 57 000) in MEDLINE.

In summary, there are a number of issues when **Get References** is chosen, and these are discussed further in Chapter 5, Section 5.2.

## 3.10 Combining Explore by Chemical Substance and Research Topic

The combination of searching for substances together with bibliographic and keyword information and the analyze/refine options are just two of the unique



and very valuable opportunities offered for searching for information in the sciences through SciFinder Scholar. It remains for the scientist to take full advantage of these opportunities.

Just one example of how the user may think creatively is illustrated through Figure 3.22. The interactions between molecules occur because of matches in shapes and the spatial positions of groups, and one way this is typified in biology is through the interactions of enzymes with their substrates. A certain amount of preorganization of molecules is necessary, and in modeling enzyme reactions, chemical scientists have turned to a variety of structural types that place groups in specific positions. One of these is the rigid structure (**49**) present in Troger's base. Not only does this molecule provide a framework on which to build, but since it is chiral, it also allows stereochemical issues to be addressed.

SciFinder Scholar allows many opportunities for scientists to explore how this structure has been used in the study of intermolecular interactions. Thus, a substructure search on (**49**) gives 476 substances, which may be analyzed

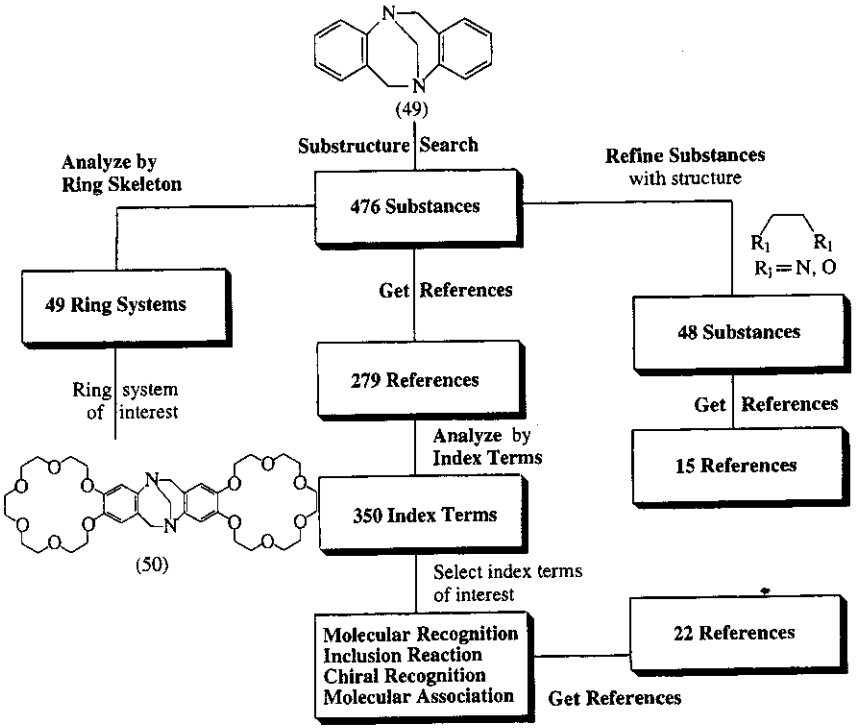



**Figure 3.22** Summary of options that combine structure and text searches.

by ring skeleton. In this case, 49 ring systems are identified including those containing various porphyrins and crown ethers (e.g. ring skeleton (**50**)). The graphic representation of the ring skeletons makes it very easy to understand exactly the types of groups that have been attached to the Troger's base framework.

Accordingly, the 476 substances may be refined with a part structure. When refined with ethylene glycol–related functionality ($R_1$–C–C–$R_1$ where $R_1$ = O,N), 48 substances are retrieved and references on them may readily be obtained.

Alternatively, references for all the original substances may be obtained, and while these may be refined with author/indexer related terms, the user may wish to explore systematic indexing through **Analyze** by **Index Term**. Selection of key index terms gives a new answer set of 22 references.

Although the user no doubt would be delighted with the results, nevertheless, it helps to think carefully about the potential limitations of the searches and about alternative strategies. Since the search commenced with a structure search, the user is relying on CAS Registry Numbers to be present in relevant answers, but the indexing of CAS Registry Numbers is highly efficient when something new is reported for the substances. So, probably, this search has nicely combined issues of precision and comprehension. Alternative strategies starting with **Explore by Research Topic** could be undertaken, and some of the issues here are addressed in Chapter 5, Section 5.2.

## 3.11 Summary

Key points relating to structure searching on SciFinder Scholar are as follows:

- The representation of chemical structures and chemical bonds is quite complex, and it helps if searchers understand some of the structure conventions used by CAS in order to obtain good answer sets.
- Particular issues relate to
  - aromatic compounds/resonance;
  - tautomers;
  - $\pi$-bonding; and
  - donor bonding.
- SciFinder Scholar automatically interprets a lot of these issues, but at times more precise answer sets need to be obtained through **Analyze Substances**, then **Precision**.



- In substructure searches, SciFinder Scholar defaults to
  - rings are isolated or fused;
  - chains are assigned chain or ring values;

  and the locking tool is used to override the defaults.
- When the user chooses **Get References** from a structure answer set, Scholar searches for the CAS Registry Numbers for the substances in CAPLUS and MEDLINE.
- Answers from bibliographic databases may be analyzed or refined, for example, by research topic, bibliographic information, or index terms.
- The user needs to think carefully about the structure to search and the type of search (exact/related or substructure) required. Will a more general search (e.g. one that may include synthetic precursors) be better?

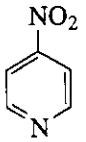
The real challenge, however, is to draw the structure that best suits the requirements! As a general principle in searching for information, it is better to start with a broader search and then refine answers rather than to start off too precisely. In this way, not only is the risk of missing key answers reduced, but related information may also be retrieved, which in fact, may prove more valuable than just that obtained through a very specific search. So, as a general rule, the user should start the search at a more general level, then if the initial answer set is too large, the user may narrow the search with options under **Refine Substances** or **Analyze Substances**, build more atoms into the structure, or just **Get References** for the large number of substances and refine/analyze with topics in the bibliographic databases.

## 3.12 Exercises

### *3.12.1 General Questions on Operation of SciFinder Scholar*

Consider the questions (3.1–3.12). Try to work out the various issues before you log in to SciFinder Scholar. Then enter Scholar and try the problems. (How well did you anticipate the issues?!)

3.1 Give alternative ways to find 4-nitropyridine (A) in the substance database. What are the potential problems for each strategy?



(A)

3.2 It was indicated that a substructure search (phenyl ring locked) on structure (**2**) (see Section 3.4) gave substances with structures (**3**) to (**7**). Why were these substances retrieved, and what strategies may be used to restrict answers to primary amines containing 2-aminomethyltetrahydrofurans with aromatic rings at the 5-position?

3.3 Which of the following structures will be retrieved when the structure query (B) is searched:

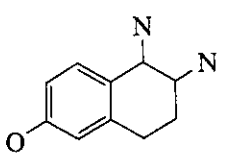
(i) in an exact/related structure search;

(ii) in a substructure search (no defaults overridden);

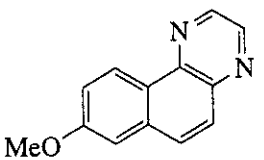
(iii) in a substructure search (ring locked);

(iv) in a substructure search (chain locked);

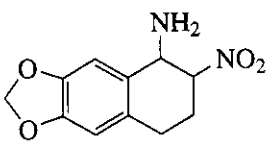
(v) in a substructure search (ring/chain locked).



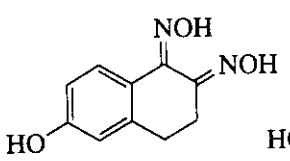
(B)



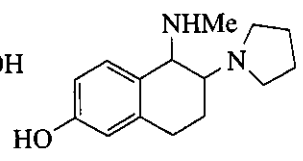
(B1)



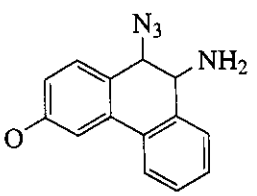
(B2)



(B3)



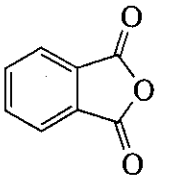
(B4)



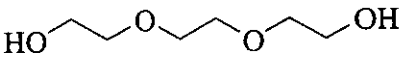
(B5)

3.4 Suggest ways to find all polymers that have the monomers phthalic anhydride (C), and triethylene glycol (D). Would the same, or different, answers arise if phthalic acid was drawn instead of the anhydride?



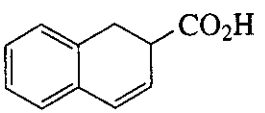


(C)

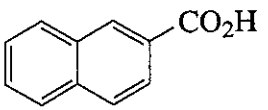


(D)

3.5 Substances with the dihydronaphthalenecarboxylic acid part structure (E) are required. The structure is drawn and a substructure search is undertaken, but a large number of the answers are 2-naphthalenecarboxylic acids (F). Why are these substances retrieved, and how may just the dihydronaphthalenes be retrieved?

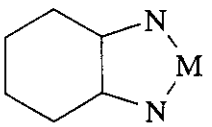


(E)



(F)

3.6 Substances of the type (G) are required where M is a metal of the periodic table group I or II. How would you find these substances?

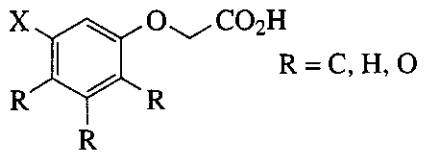


(G)

M = Li, Na, K, Rb, Be, Mg, Ca, Ba

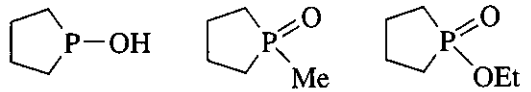
3.7 Information on the preparation of xylenes (e.g. from refining of petroleum products) is required. It is known that in the database there are three specific xylenes (*o*-, *m*-, and *p*-xylene) as well as the incompletely defined substance (Appendix 5.5.1). Propose a strategy to obtain all these substances in a single answer set in the substance database.

3.8 Tri-substituted benzenes with the part structure (H) are required, but when a substructure search is performed, di-, tri-, tetra-, and penta-substituted benzenes are retrieved. Why is this so, and what is the best way to retrieve only the tri-substituted benzenes required?



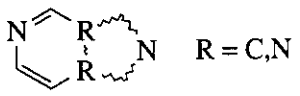

(H)

3.9 Substances of the type (I) are required, but a substructure search on structure (I) produces more than 2000 answers, which include substances (J) and (K). Why are these compounds retrieved, and what would be a good way to select out just those compounds of the type (I)?



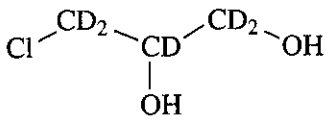

(I) (J) (K)

3.10 Substances of the type (L) are required in which at least one of the atoms R must be a nitrogen. However, a search on (L) produces many compounds that have both Rs = C. Suggest a solution to retrieve compounds in which at least one of the groups R is a nitrogen.




(L)

3.11 How would you find the substance (M) (including the stereoisomers)?




(M)

3.12 Compounds of the type (N) are required. However, when this structure is drawn and searched by substructure, more than 10 000 substances are retrieved. (Remember that the bond to the metal is ignored in the initial search!) The usual procedure now is to **Analyze Substances** by **Precision** (but this may take some time because of the large number of



answers to look through). So, try a search on structure (O), and then **Analyze Substances** by **Variable group composition**. Comment on the two processes and in particular on why in the second search, the initial answer set is much less than 200 substances.


(N) (O)

### *3.12.2 Additional On-Line Questions*

Find the following substances. In all cases, note the entries in the name, formula, and structure fields, and thus consider whether alternative methods to find the substances may have been possible.

3.13 Cyclosporin A

3.14 Stanozolol

3.15 Anhydrous calcium carbonate

3.16 Sildenafil and all it salts

3.17 Propylene glycol (including the specific stereoisomers)

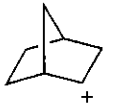
3.18 Augmentin

3.19 Clavulanic acid and all its salts

3.20 All dichloronaphthalenes (including isotopically labeled substances)

3.21 Polystyrene (homopolymer)

3.22 The norbornyl carbocation (P)




(P)

3.23 The copolymer of styrene, 1,3-butadiene, and acrylonitrile

3.24 The CAS Registry Number for $CH_5^+$ (the protonated form of methane)

3.25 Titanium aluminum nitride

Chapter 4

# Additional Search and Display Options

## 4.1 Introduction

SciFinder Scholar offers many additional features. Some of these are explore options (Author Name, Accession Number, Company Name, and Sequence Searching) that afford initial answer sets, while others (Get Related and Property Information) are functions that are used once the initial answers are obtained or are data mining and visualization tools (Panorama and Spotfire® DecisionSite). Two other functions are Browse Table of Contents and Keep Me Posted (current awareness searching). At present, some of these features are available only in SciFinder, and additionally, access to Spotfire DecisionSite depends on users having independent arrangements with the software provider.

Key aspects of these additional features are discussed in this chapter.

## 4.2 Explore by Author Name

The exact entry of an author name in an original scientific article depends partly on how the authors wrote their names in the article and partly on the policies of the editor of the publication. Variations in representations occur mainly with entries for first names when either full first names or just initials appear. The situation is further complicated by the database producers who may apply additional policies.

For example, it is clear from Figure 1.5 that MEDLINE uses only initials, whereas CAPLUS (Figure 1.3) uses exactly the entry in the original article.

Complications may also occur through the ways in which databases interpret phonetics (e.g. Müller is entered as Mueller), prefixes (e.g. van and de), and hyphenations. Names may be translated into American English (e.g. most Chinese-language documents have author names as Chinese characters, and names that appear in the database are translated). Further, in some original documents, the author's last and first names may be in different word orders, and this may cause complications for the indexer. Names may also be misspelled. Consequently, a single author may be represented in many ways, for example, Professor Erik de Clercq appears in more than 20 different ways in the two databases (Section 2.5.4 and Figure 2.9 of Chapter 2).

In SciFinder Scholar, authors are searched through **Explore by Author Name** (Figure 4.1). By default, the list of candidates includes variations in spellings in last names but not in first names, and because of the latter, it is advisable to enter initials only. If initials are entered, then candidates with possible full first names appear, whereas if full first names are entered, candidates with the corresponding initials appear. Because of the possibility of many alternatives, the entry should be kept as short as possible (e.g. last name and first initial), although if the name and initial are quite common, the list of candidates may be very extensive and may take some time to work through.

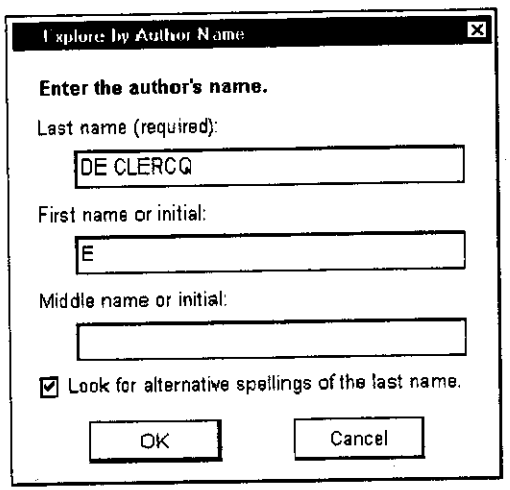



**Figure 4.1** Screen for **Explore by Author Name**. Note that alternative spellings of the Last name but not of the First name are searched, so it is advisable to enter initials only. SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.



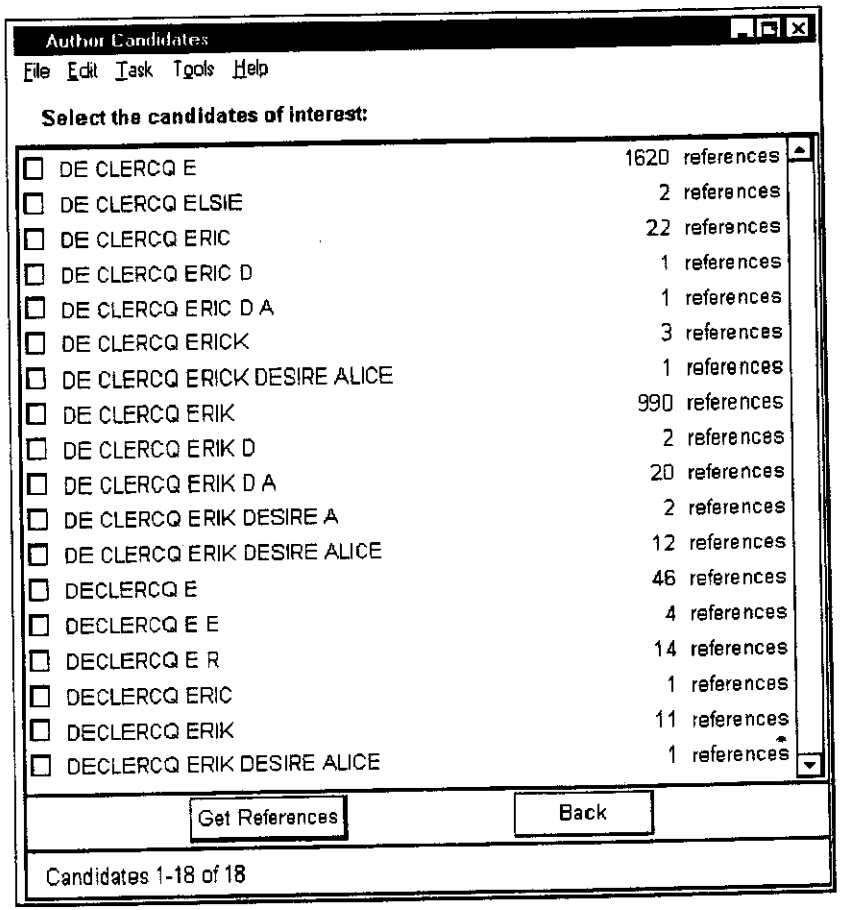



**Figure 4.2** Candidates obtained from search in Figure 4.1. Note alternative spellings of De Clercq, and potential full first names and initials are retrieved. If first name 'erik' had been entered in Figure 4.1, then candidates with first name 'eric' would not be retrieved. SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

The entry in Figure 4.1 produces 18 candidates (Figure 4.2) and indicates some alternative spellings (e.g. Declercq) for the last name. The algorithm did not pick up entries for De Klerk, but to use an algorithm that allows for such variations would be very difficult and in any case, depending on the author, many false hits may result. If the entry in Figure 4.1 is 'Last name "Declercq"', it is found that the first 12 candidates in Figure 4.2 are not shown. Again, it is a matter of the sophistication of the algorithm. The point is that although SciFinder Scholar helps a great deal, if absolutely

comprehensive results are required, the user still may have to use a few alternatives to ensure that all records are identified.

The other major issue relating to searching for authors is that different authors may have the same names, or at least, the same last names and initials. So, while Margaret M Harding (University of Sydney) and Marjorie M Harding (University of Edinburgh) may be distinguished by their first names, they may also be represented as M M Harding in original documents. Further, MEDLINE will represent them only as M M Harding. The usual way to resolve this issue is to add extra search terms to answer sets, for example, to additionally search by company/organization or by research topic. However, it should be remembered that both CAPLUS and MEDLINE list only one affiliation (so, in the case of multiauthored papers, the Universities of Sydney or Edinburgh may not be listed), and in this particular case, both authors share at least some research interests (in protein chemistry). The problem is often more complex for some names of Chinese, Korean, or Taiwanese origin, where many authors share the same last name and may have only one first name (which again is used by many different scientists).

How crucial these issues are depends on the intent of the search and on how unusual the name is. If the intention is to find the majority of papers written by an author, then an author search followed by refinements is usually sufficient. If the intention is to get a complete list of papers for an author, then all possible variations of the author's name must be considered, and it is usually necessary to look through all the records to check that papers from different authors are not included.

## 4.3 Explore by Accession or Patent Number

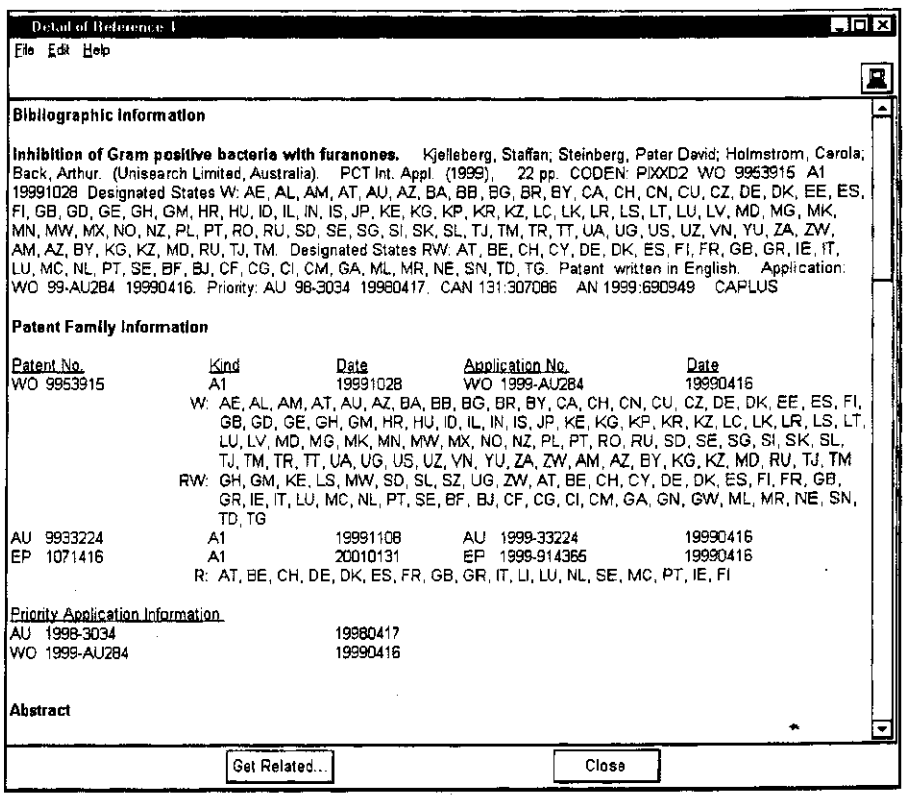
Accession numbers are used by database producers as unique identifiers for records and are found in the electronic database (or the print equivalent). If the searcher wishes to retrieve a record, for example, to display the full record, to check CAS Registry Numbers, or to work through citation links or full text, then the accession number may be entered directly if it is known from a previous search.

Patent numbers are used by patent authorities as unique identifiers for patents, but many numbers may relate to the single invention as it passes through the patenting process. For example, a *patent application number* is assigned when the manuscript is first received, and a *patent number* is assigned when the manuscript has passed the full examination process. Patents need to be taken out in all countries where the inventor seeks coverage, and so a single invention may have many application and patent



numbers. All these numbers make up what is known as the patent family information (Figure 4.3), and searches on any of the numbers will retrieve the record.

Scientists may need to consult information professionals for a full explanation of the various patent numbers and codes, although users of SciFinder Scholar generally are interested only in the scientific aspects of the document. Indeed, questions related to the legal status of patents most probably need to be asked only if the searcher is interested in the patentability of their research, and such questions are not within the usage restrictions for SciFinder Scholar. However, finding scientific information in patents is an important issue for academic staff and students and is within the usage restrictions for SciFinder Scholar.


Detail of Reference 1

File Edit Help

**Bibliographic Information**

**Inhibition of Gram positive bacteria with furanones.** Kjelleberg, Staffan; Steinberg, Peter David; Holmstrom, Carola; Back, Arthur. (Unisearch Limited, Australia). PCT Int. Appl (1999), 22 pp. CODEN: PIXXD2 WO 9953915 A1 19991028 Designated States W: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZA, ZW, AM, AZ, BY, KG, KZ, MD, RU, TJ, TM. Designated States RW: AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, BF, BJ, CF, CG, CI, CM, GA, ML, MR, NE, SN, TD, TG. Patent written in English. Application: WO 99-AU284 19990416. Priority: AU 98-3034 19980417. CAN 131:307086 AN 1999:690949 CAPLUS

**Patent Family Information**

| Patent No. | Kind | Date | Application No. | Date |
|---|---|---|---|---|
| WO 9953915 | A1 | 19991028 | WO 1999-AU284 | 19990416 |
| | W: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZA, ZW, AM, AZ, BY, KG, KZ, MD, RU, TJ, TM | | | |
| | RW: GH, GM, KE, LS, MW, SD, SL, SZ, UG, ZW, AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG | | | |
| AU 9933224 | A1 | 19991108 | AU 1999-33224 | 19990416 |
| EP 1071416 | A1 | 20010131 | EP 1999-914365 | 19990416 |
| | R: AT, BE, CH, DE, DK, ES, FR, GB, GR, IT, LI, LU, NL, SE, MC, PT, IE, FI | | | |

Priority Application Information

| | |
|---|---|
| AU 1998-3034 | 19980417 |
| WO 1999-AU284 | 19990416 |

**Abstract**

Get Related... Close

**Figure 4.3** Detail of a reference showing patent family information. The patent numbers, priority application numbers, and countries covered are important legal aspects of the document. Copyright the American Chemical Society, and reproduced with permission.

SciFinder Scholar contains records for more than three million patents and information on the patents covered is available through *http://www.cas.org/EO/caspat.html*. The majority of these records are linked to full text documents. Records have titles, abstracts,[1] and full indexing, and the citations within patents occur from 1999 onward; so, SciFinder Scholar provides many entries into the patent literature at the desktop.

## 4.4 Explore by Company Name/Organization

Searching for company names is a very complex task, and some issues that need to be kept in mind include the following:

(1) authors may state their organization in a number of ways (e.g. compare the entries for the Rega Institute in Figure 2.10);

(2) CAPLUS and MEDLINE may have file-specific abbreviations (e.g. compare the affiliations listed in Figures 1.3 and 1.5);

(3) company mergers or acquisitions may lead to changes in company names from time to time; and

(4) usually, only one affiliation is listed in CAPLUS and MEDLINE, even though separate affiliations for different authors may be given in the original article (e.g. the original article represented in Figures 1.3 and 1.5 also contains the affiliation for Jonathan Sessler at the University of Texas).

Information on organizations may be found in two different ways. First, a bibliographic answer set may be obtained (e.g. by **Explore by Research Topic**) and subsequently, the answers may either be analyzed or refined by company name/organization. Alternatively, **Explore by Company Name/Organization** may be used.

The advantage of **Analyze References** by **Company Name/Organization** is that individual listings and variations are indicated in the histogram and may be selected. The limitations are that the list of options may be very long and that the initial answer set may not have retrieved all the relevant records. Nevertheless, in this way, a reasonable indication of the types of entries present may be obtained, and if more general searches through **Explore by Company Name/Organization** are required, then strategies may be developed on the basis of the terms identified.

[1] Note that titles and abstracts in records for patents may be rewritten by the indexers when the original document does not contain sufficient technical information in these fields.



On the other hand, both **Refine References** by **Company Name/Organization** and **Explore by Company Name/Organization** require precise inputs, and as already discussed, this needs to be done with caution because of the variations of the entries in the databases.

At present, in SciFinder Scholar, there is no way around the restriction that only one affiliation is listed in the databases. When comprehensive information on the company or organization is required, alternative databases that list all affiliations in the original documents may need to be consulted. However, all databases are restricted through the points 1 to 3 listed above, and caution needs to be exercised in interpreting the results.

**Explore by Company Name/Organization** is used when the search needs to be started through the name of an organization (Figure 4.4). In the search, SciFinder Scholar uses a number of algorithms to identify possible candidates, including searches on the complete company name, on individual words in the company name, and on its company dictionary. This dictionary of company synonyms means that alternative names for the company may be searched in addition to the names entered.

Generally, it is preferable to enter as few terms as possible, and of course, to choose terms that may be relatively unique. Accordingly, if records for articles from the Rega Institute (Figure 2.10) are required, then the entry in Figure 4.4 gives a set of more than 2200 records of which two-thirds are from CAPLUS. These records may then be analyzed in the usual way, and some of the most listed terms are shown in Figure 4.5.

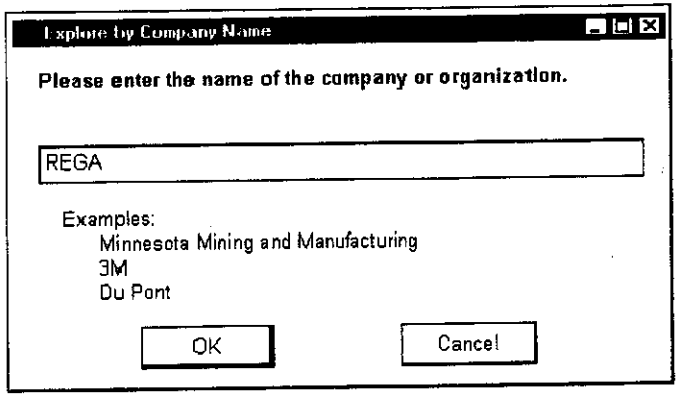



**Figure 4.4** Screen for **Explore by Company Name**. It is preferable to enter as few terms as possible in the first instance and then to analyze/refine initial answer sets. SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

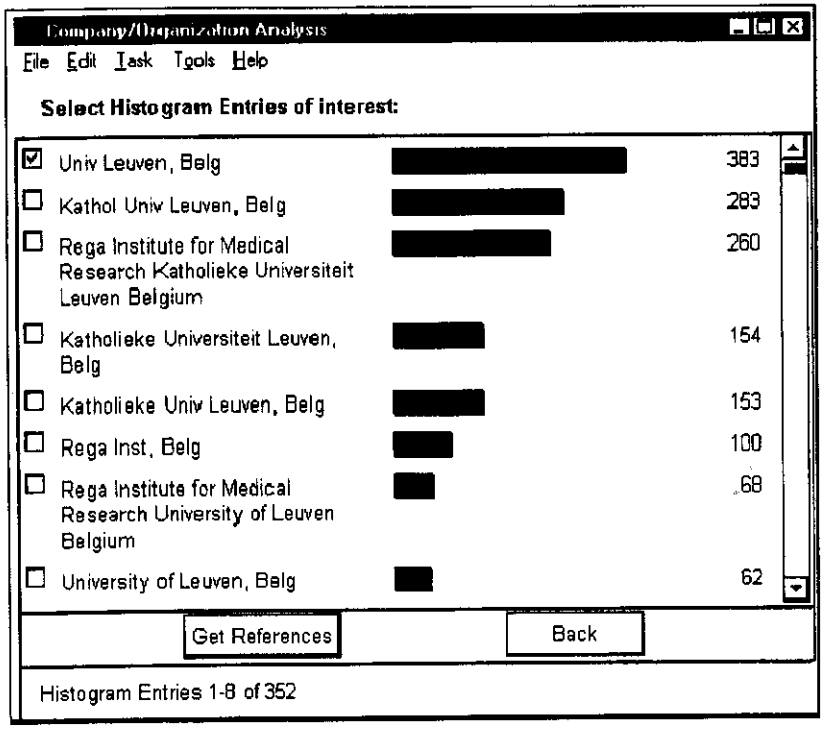



**Figure 4.5** Part of the histogram of entries when the result of the search in Figure 4.4 is analyzed by Company Name. For many company listings, particularly for universities, entries may include the section or department as well as the parent company. The histogram may include hits of either or both types, although when records are retrieved, they will contain the full affiliations. SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

Names of institutions may be very lengthy, so the full name of the institution may not be shown in the histogram. Additionally, the search field may contain subfields with information on the department, the institution, and the country, and depending on the search, different sections may be displayed in the histogram. For example, some of the actual records listed under the first row in Figure 4.5 ('Univ Leuven, Belg') have the *full entry* 'Rega Institute for Medical Research, Katholieke Univ. Leuven, Louvain, Belg', and so it is often found that the number of references obtained when the various boxes are checked are much greater than those indicated in the histogram. (When Univ Leuven, Belg (first row in Figure 4.5) is checked and **Get References** is clicked, more than 800 answers are retrieved.) Accordingly, in practice, it



is necessary to check all the relevant boxes, and then to work systematically through the next set of answers.

If two or more terms are entered in the query, SciFinder Scholar looks for these terms anywhere in the field for the company/organization and depending on the specific entries in the company dictionary, may or may not add synonyms to the search. For example, if '3M' is entered, a number of synonyms including 'Minnesota Mining and Manufacturing' will be searched since these are in the dictionary. However, if '3M Japan' is entered, the algorithm may retrieve records with the terms entered only and the synonyms may not be applied. For this reason, it is usually better to take the search in steps, and, for example, a search on '3M' (where the synonym dictionary is applied), followed by a refine with 'Japan' gives more than double the number of answers in comparison to an initial search just on '3M Japan'.

Word order is not important and searches for 'univ sydney', 'university sydney', 'sydney university', and 'sydney univ' produce identical answer sets. However, these records will contain hits for several of the universities in Sydney including Macquarie University, University of Western Sydney, and University of New South Wales, as well as The University of Sydney.

In summary, searching for Company Name/Organization presents many challenges for the searcher and for the algorithms used by SciFinder Scholar. Although the algorithms work very well in most cases, if comprehensive and precise answers are required, it is probably better to refer the problem to an information professional.

## 4.5 Getting Related Information

Once a bibliographic answer set has been obtained, and after the **Get Related** button at the bottom of the screen has been clicked, a new screen appears (Figure 4.6). Presently, the third option, **Substances**, is available only to SciFinder Scholar and SciFinder subscription users only.

### *4.5.1 Citations*

Citations have been added to the CAPLUS database from the beginning of 1999, and so cover many publications from 1998. Further information on the coverage may be obtained through *http://www.cas.org/STNEWS/MAYJUN00/feature.html*, and here, it will be noted that a unique aspect is the coverage

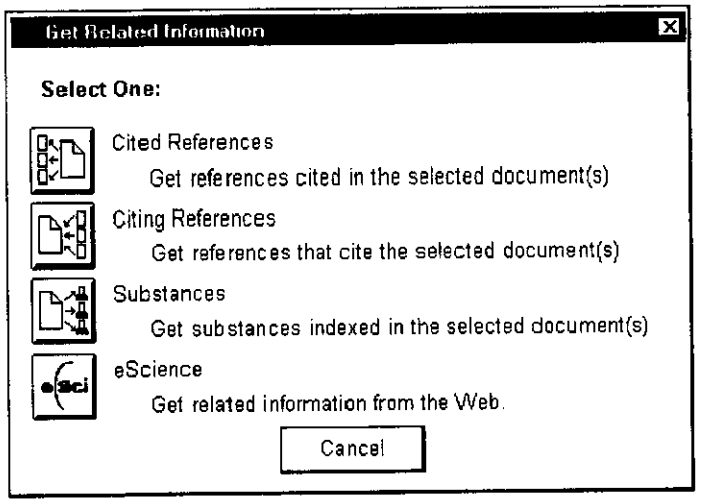



**Figure 4.6** Options available under **Get Related** (e.g. see Figure 2.4). This screen appears for SciFinder Scholar and SciFinder subscription users. Other users do not have the **Substances** option. SciFinder screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

of citations from patents. While there are a number of issues relating to citations,[2] nevertheless, citation linking and citation searching provide further opportunities for information retrieval.

As answer sets that involve cited or citing references can quickly become very large, the ability to work through larger answer sets in systematic ways with the many options available in SciFinder Scholar is very important.

### Cited References (Working Backward in Time)

Cited references are displayed automatically as part of the record, and Figure 4.7 gives the first few of the 38 citations in the complete record. Links to records in the database are provided for each citation, but when links are not indicated, it simply means that a record for the citation is not present. This often occurs for books (e.g. Citations 1 and 2 in Figure 4.7) or because the algorithms used to link citations may not have recognized the specific page numbers mentioned in the original citation.

Although the citations may be viewed in turn through the links, often it is better to view citations through **Get Related**, followed by **Cited References**,

[2] For example, see 'Citation searches in on-line databases: possibilities and pitfalls' in *Trends in Analytical Chemistry*, **20**, 2001, 1–10.



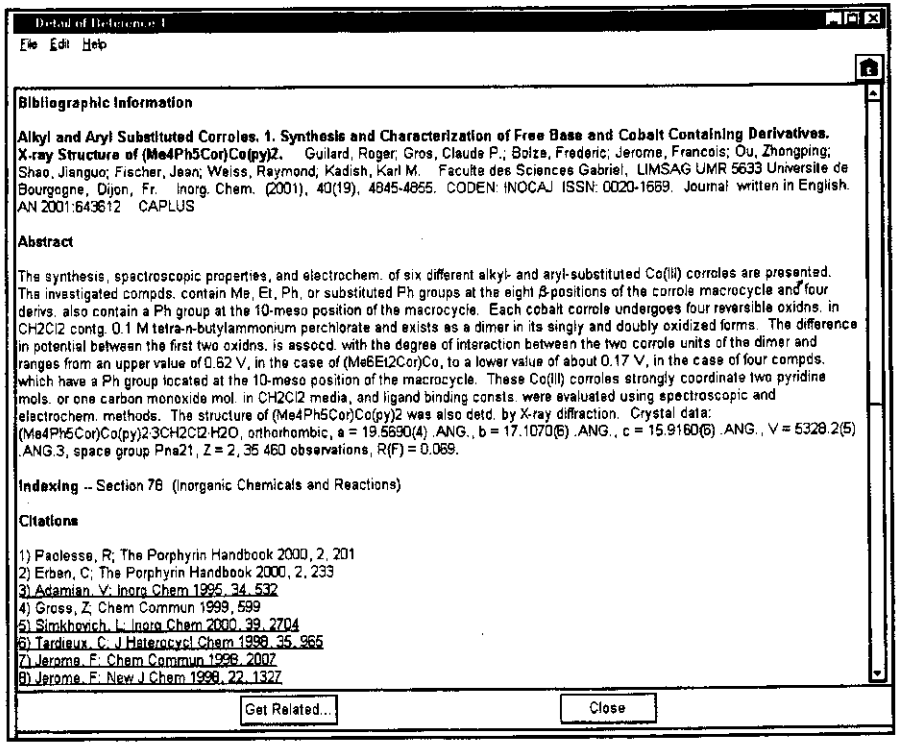



**Figure 4.7** Screen showing citations in CAPLUS. Note that this is an example of a recently added record, and that full indexing will soon appear in CAS Section 78 (Inorganic Chemicals and Reactions). Copyright the American Chemical Society and reproduced with permission.

since this process gives the complete list of records as a separate answer set. This answer set may be analyzed or refined in the usual manner, and this is often important because at times, many citations may refer to experimental details (such as melting points) and not to the subject matter, which usually is the intention in finding related records.

Accordingly, an **Explore by Research Topic** 'I am interested in ''urocanic acid with cancer'' ' gives just over 90 records in which the two concepts are 'anywhere in the reference' and **Get Related** followed by **Cited References** gives over 450 records. Since cited references are automatically linked in database records, this answer set will contain records from CAPLUS and MEDLINE.

Advantage may now be taken of the analyze/refine options, and, for example, **Analyze References** by **Index Term** indicates important headings such as 'ultraviolet rays', 'UV B radiation', and 'radiation effects', which

may be used to focus the answers and to eliminate those citations that do not relate to the topic of interest.[3]

**Citing References (Working Forward in Time)**

Working with citing references enables a searcher to investigate more recent information on the topic. Thus, after the records that contain the concepts 'urocanic acid' and 'cancer' are retrieved, and **Get Related** followed by **Citing References** is chosen, a new answer set of around 160 records is obtained, and the majority of these give more recent information on the subject. As citations are presently included only in CAPLUS, these citing references are from CAPLUS only.

Expanding searches through citations is another way to work through the issues of comprehension and precision in searching. It is simply another tool that is available, but it relies ultimately on the referencing used by the authors in the original publications. At times, it turns up important additional articles that may not have been retrieved even by a variety of searches on the topics.

Another particularly valuable application of citations is to locate citations for individual authors. First, all records published by the author are found (Section 4.2) and then the **Get Related** button is clicked. After **Citing References** is chosen, the required citations are presented. Since there are some restrictions on the number of answers that can be processed at one time, it may be necessary to break the initial answer set of publications into subsets, and this may be easily done through **Analyze References** by **Publication Year**.

As with all searches, it is necessary to understand the limitations, and here the main considerations are the accuracy with which the original search may be conducted (e.g. note that it may be difficult not only to get all the publications by the author but also not have false answers because of problems with different authors being represented by the same name terms), the years for which citations are entered in the database, and the identification of a record in the database for the citing reference. There are many issues relating to this last aspect including the accuracy of the original citation (e.g. whether the reference page number is correct) and the function of the algorithm, so the number of records obtained through **Citing References** should be considered as a guide only.

[3] Of course, scientists familiar with the field are aware that the UV absorption properties of urocanic acid has implications relating to skin cancer.



### *4.5.2 Substances*

It has already been shown (Section 2.5.4 of Chapter 2) that it is possible to **Analyze References** by **CAS Registry Number** and that this affords a histogram of the number of records that contain the various CAS Registry Numbers in the answer set. Having such histograms are valuable in their own right since they give indications of the more frequently listed (and hence, possibly more important) substances that relate to the topic.

However, if particular types of substances are of interest, this process requires the searcher to look through the various substances in turn. For example, **Explore by Research Topic** 'I am interested in ''traditional Chinese medicine'' ' gives more than 6500 records that contain more than 3400 substances, and the most commonly indexed substances are simple atoms such as calcium. So, how could the more complex organic substances be identified in this large list?

The answer is to **Get Related** information on **Substances** (Figure 4.6; this function is presently available only for SciFinder Scholar and SciFinder subscription users) and as this feature is available for answer sets of less than 1000 records, in this case, subsets of the original answers need to be obtained. In the event, after answers from an explore on 'traditional Chinese medicine' are restricted to English language in CAPLUS, and after **Get Related** and then **Substances** is chosen, more than 1700 substances are displayed in REGISTRY. The searcher may now use the usual substance refine/analyze options, and indeed, an analysis of nitrogen-containing ring systems provides interesting answers (Figure 4.8).

This option fully integrates all the unique functionalities for processing answers in CAPLUS/MEDLINE with those in REGISTRY. Thus, at any stage, the substances in an answer set obtained through a bibliographic search may themselves be analyzed or refined. References may be obtained for the new substance answer set, and the process may be repeated. In Chapter 6, it will be seen how the reaction database in turn may also be fully integrated.

### *4.5.3 eScience*

In a manner similar to that through which ChemPort (Section 2.5.2 of Chapter 2) provides the link from SciFinder Scholar to full text information from publishers through the Web, eScience *http://escience.cas.org* provides the link from SciFinder Scholar to Web search engines. The link is activated through **Get Related** then **eScience** (Figure 4.6), and the opening screen is shown in Figure 4.9. From this point, any of the usual Web search functions may be implemented, and in particular, immediate access to the Web search engines powered by ChemIndustry, Google, and ChemGuide is available.

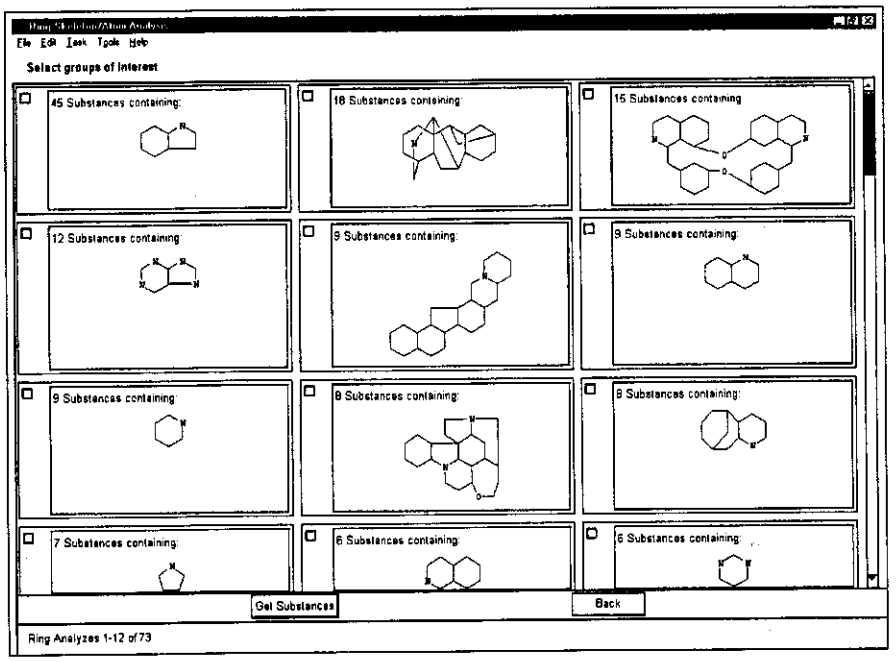



**Figure 4.8** Example of analysis by **Ring Skeleton with atoms** for an answer set of substances obtained from CAPLUS records through **Get Related** and then **Substances**. SciFinder screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

When eScience is activated after searches initiated through **Explore by Research Topic, Author Name**, or **Company Name/Organization**, the concepts in the original SciFinder Scholar query are directly entered in the search box, although prepositions/conjunctions and certain controlled terms are not transmitted. The text box may be edited if desired so that the functionality of the search engine (e.g. Boolean operators, truncation, and quotation marks/phrase searches) may be utilized.

So, **Get Related** then **eScience** from answers from an initial search 'I am interested in ''urocanic acid with cancer'' ' automatically enters terms as shown in Figure 4.9. Now, it is important that the user understands the search strategies available through the search engine, and, for example, in the case of a search through ChemIndustry, the query effectively looks by default for *any* of the terms. As there are numerous records that simply refer to 'cancer' or 'acid', the user is better off searching just for 'urocanic' when around 70 hits are obtained. The ChemIndustry search on 'urocanic acid cancer'



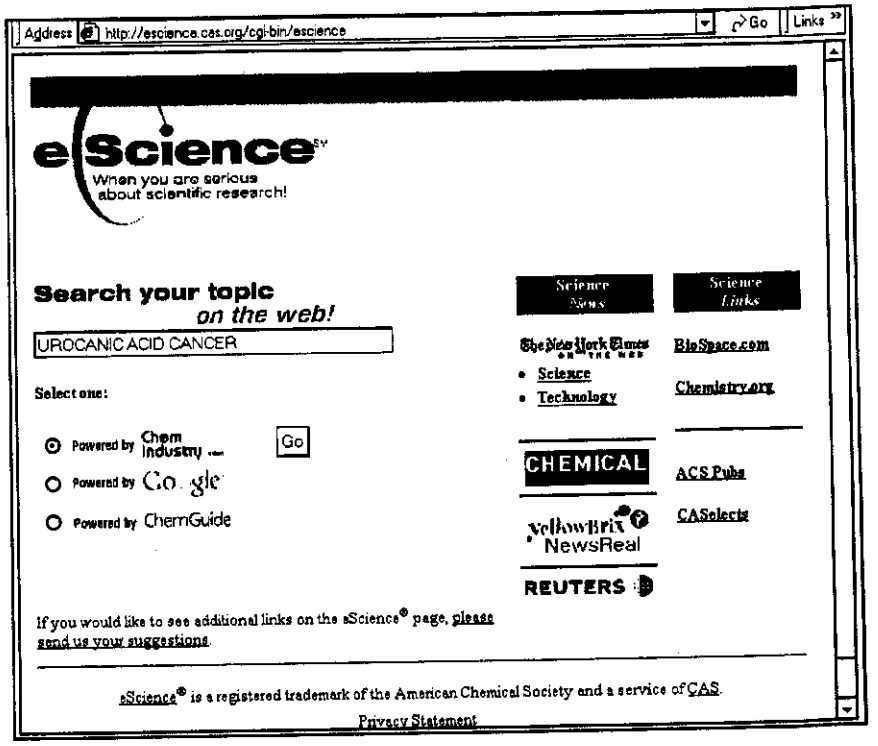


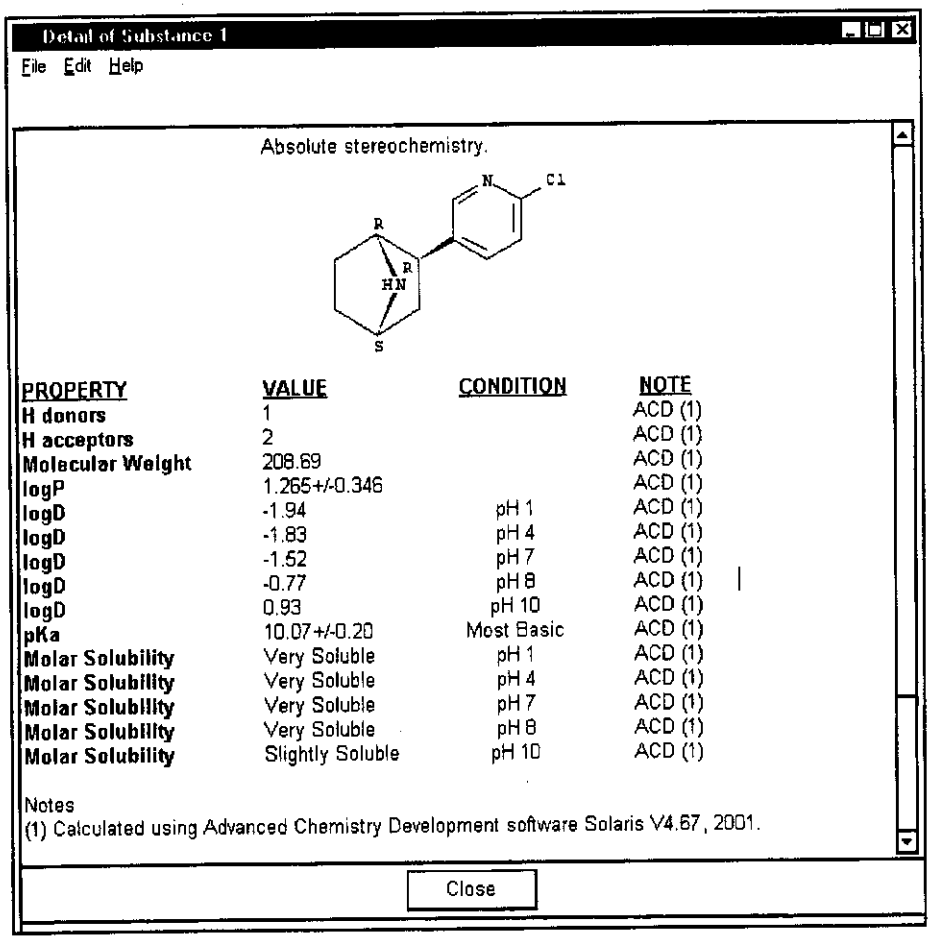

**Figure 4.9** Screen from **Get Related** then **eScience** from which ChemIndustry, Google, or ChemGuide searches may be conducted. This screen is reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

gives more than 350 000 hits and that on 'urocanic acid' gives more than 200 000 hits.

As usual, it is helpful, even with Web search engines, to understand how the search is being conducted, to understand basic search functionalities, and then to think carefully about the answer sets produced. For example, when an **Explore by Author Name**: Kadish K is attempted in SciFinder Scholar, the searcher would naturally check the boxes next to candidates: Kadish Karl, Kadish Karl M, and Kadish K M. If after the initial answers are obtained the search is extended into eScience (e.g. to find more information on the author including a Web address), then SciFinder Scholar enters all the three name terms above, and the three different Web search engines interpret the query quite differently. Therefore, the text automatically entered usually needs to be edited in such a way that optimal results are obtained for the different search engines.

## 4.6 Property Information

Calculated physical property information (*http://www.acdlabs.com*) and experimental property values are available in REGISTRY, and an example of a record is given in Figure 4.10. The sources of the information are varied, but a recent summary may be found at *http://www.cas.org/ONLINE/STN/STNOTES/stnotes27.pfd*. The inclusion of hydrogen-bonding donor and




| PROPERTY | VALUE | CONDITION | NOTE |
|---|---|---|---|
| H donors | 1 | | ACD (1) |
| H acceptors | 2 | | ACD (1) |
| Molecular Weight | 208.69 | | ACD (1) |
| logP | 1.265+/-0.346 | | ACD (1) |
| logD | -1.94 | pH 1 | ACD (1) |
| logD | -1.83 | pH 4 | ACD (1) |
| logD | -1.52 | pH 7 | ACD (1) |
| logD | -0.77 | pH 8 | ACD (1) |
| logD | 0.93 | pH 10 | ACD (1) |
| pKa | 10.07+/-0.20 | Most Basic | ACD (1) |
| Molar Solubility | Very Soluble | pH 1 | ACD (1) |
| Molar Solubility | Very Soluble | pH 4 | ACD (1) |
| Molar Solubility | Very Soluble | pH 7 | ACD (1) |
| Molar Solubility | Very Soluble | pH 8 | ACD (1) |
| Molar Solubility | Slightly Soluble | pH 10 | ACD (1) |

Notes
(1) Calculated using Advanced Chemistry Development software Solaris V4.67, 2001.

Close

**Figure 4.10** Structure and property display in the record for epibatidine (CAS Registry Number 140111-52-0) in REGISTRY. Copyright the American Chemical Society and reproduced with permission.



acceptor data and of acid dissociation constants ($pK_a$) are of particular relevance to studies on intermolecular associations. Meanwhile, log $P$ and log $D$, the partition coefficients in octanol/water mixtures for the neutral species and the actual species at various pH values, respectively, give valuable information on solubility in biological media. Information on the calculation and reliability of the data may be obtained through the ACD Web site, particularly, *http://www.acdlabs.com/products/phys_chem_lab/logd* and *http://www.acdlabs.com/products/phys_chem_lab/pka/batch.html*.[4]

As an example of an effective way to use property information, consider a query that aims to identify pharmacophores for epibatidine (Figure 4.10). First, it is necessary to define the key part of the structure of interest and to perform a substructure search (e.g. of the substructure in Figure 4.11).

This gives more than 10 000 substances, and when **Refine Substances** and then **Property Data** are chosen, the relevant data is entered into the screen (Figure 4.12). It usually is better to make the restrictions in stages, since in this way the user may evaluate the implications of the alternatives in turn. Therefore, as a first option, substances with 'no value' (i.e. those that currently do not have property values listed) may be excluded, and the answer set now contains around 3500 substances that may be further analyzed or refined. From here, there are eight different calculated property values to consider, including the two extra options under **Refine Substances** (i.e. by structure or by commercial availability). For example, the user may refine the substances with a query in which the chain bonds in Figure 4.11 are locked and then may explore outcomes when different numbers of 'freely rotatable bonds' are searched. At every stage in the process, three-dimensional representations of specific substances may be viewed through the associated link to **3D Model**.

Answers obtained in this manner give an excellent idea of the substances reported in the literature for a particular substructure, of how they look in three-dimension, and of some of their key properties. Log $P$, log $D$, $pK_a$, solubility data, and the number of freely rotatable bonds provide valuable information to the medicinal chemist, although it should be remembered that the conformational issues inherent in the bond rotations should always be considered along with the energy surface values for the docking molecule. Of course, the *spatial positions* of the hydrogen donors and acceptors and

[4] For further information on how to interpret and apply the data, perhaps start from the review article by Christopher Lipinski. The SciFinder Scholar record for this review is easily obtained through **Explore by Document Identifier:** 2001:207214.

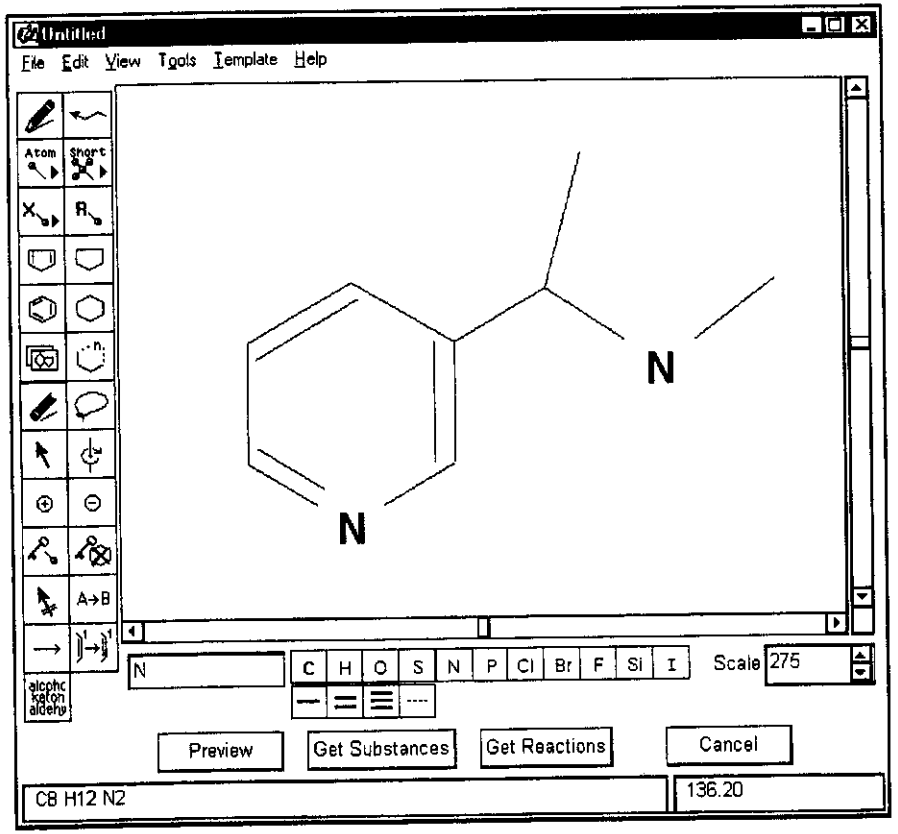



**Figure 4.11** Screen for structure search for pharmacophores for epibatidine. SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

issues of $\pi$-stacking are other extremely important aspects of intermolecular interactions. Here, SciFinder Scholar helps through searches involving the number of donors and acceptors, and answers may be refined by queries involving aromatic rings to give indications of $\pi$-stacking groups. SciFinder Scholar thus provides quite specific lead compounds for the more extensive computer analyses needed to fully understand molecular associations.

## 4.7 Sequence Searching

Protein and nucleic acid sequence searches through the Basic Local Alignment Search Tool (BLAST®) are now available to SciFinder users. BLAST



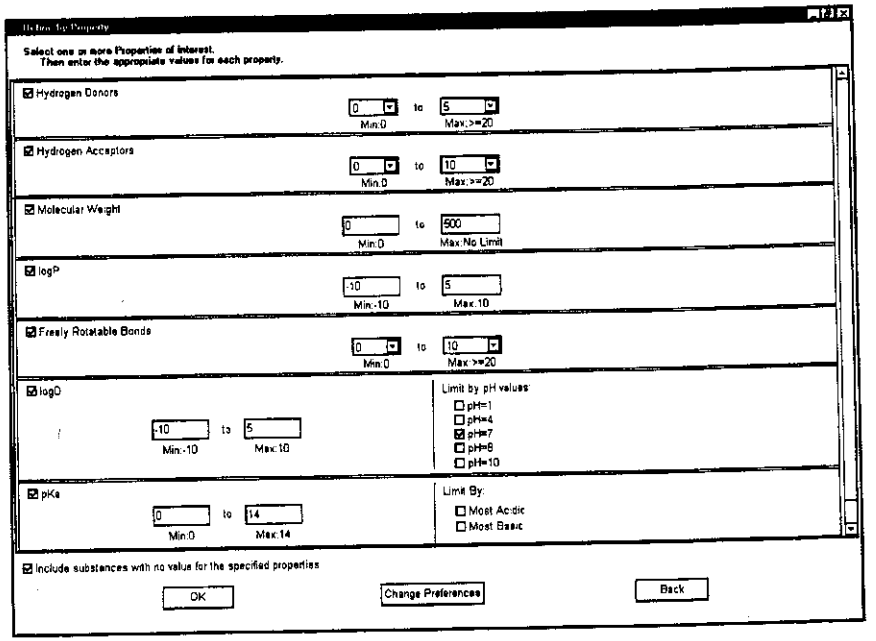



**Figure 4.12** Screen for input of property information. (Note: seven of the property fields are displayed, the remaining field is 'solubility'.) SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

programs are a suite of sequence comparison algorithms, and general information on the algorithms and the concepts involved are available in many places on the Web (e.g. *http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/similarity.html*). More specific information on the application of BLAST algorithms may be retrieved through SciFinder Scholar 'I am interested in "BLAST"' or by looking at citations to the original article by Stephen Altschul *et al.* (see Exercise 4.7).

BLAST programs are used to search sequence databases, and in the case of SciFinder, they are used to search the REGISTRY database. The search commences from the icon labeled **Nucleotide or Protein Sequence** on the SciFinder main screen (Figure 4.13), and after the sequence is entered, the appropriate BLAST Program is selected (Figure 4.14).

Figure 4.15 presents an example of answers obtained through a typical nucleic acid similarity search, and many BLAST display options are available through the various icons and check boxes. Of particular note is that answers are from REGISTRY and the full REGISTRY details may be viewed by clicking on a substance link.

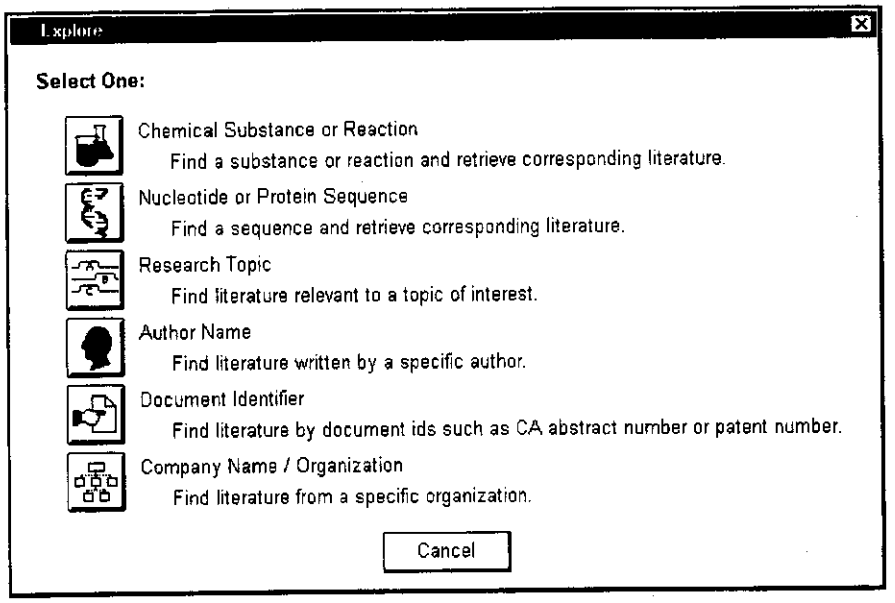



**Figure 4.13** SciFinder Explore screen. SciFinder screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

Finally, when **Get References** (Figure 4.15) is chosen, the screen shown in Figure 3.21 appears, and from here, bibliographic records from CAPLUS and MEDLINE may be obtained.

Since BLAST is available through other systems, the scientist needs to consider whether sequence searches should be performed through SciFinder in the first place, and there are two main points to note. First, REGISTRY is probably the largest single source of protein and nucleic acid sequences. It contains more than 1.6 million protein sequences and more than 21.6 million nucleic acid sequences including over 18.9 million nucleic acid sequences in GENBANK® and over 2.7 million additional sequences that have been indexed from CAS sources.[5] These *additional* nucleic acid sequences are reported in more than 140 000 records in CAPLUS, and almost half of these are records from patent documents. Indeed, the number of sequences unique to patent documents are growing so rapidly, and are of such great importance, that no search is comprehensive unless these are also covered. The other

[5] The figures quoted here are current to mid-October 2002. However, sequence databases are growing very rapidly and, if it is critical to determine the current number of sequences, the searcher should consult database producers for latest figures.



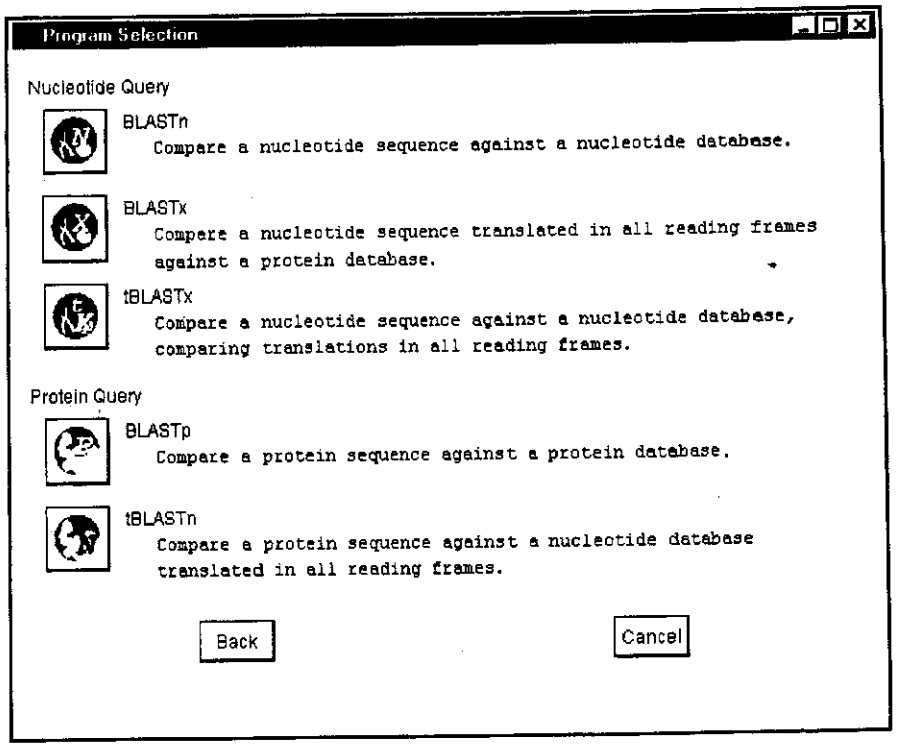



**Figure 4.14** BLAST® programs available within SciFinder and their functions. SciFinder screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

major source of sequences from patents is the DGENE database on STN *http://www.cas.org/ON-LINE/DBSS/dgeness.html*.

Second, while answers from BLAST searches in SciFinder are substances, the corresponding references in CAPLUS and in MEDLINE may easily be obtained and the references may be further processed through any one of the unique SciFinder explore/analyze/refine options.

## 4.8 Data Mining and Visualization

As databases increase in size, users obtain greater number of records, and the analyze/refine options within SciFinder Scholar help greatly to narrow down answers in a simple and predictable manner. However, in recognition of the complexities of viewing data from large databases, SciFinder now additionally offers a data mining option called *Panorama*, and users

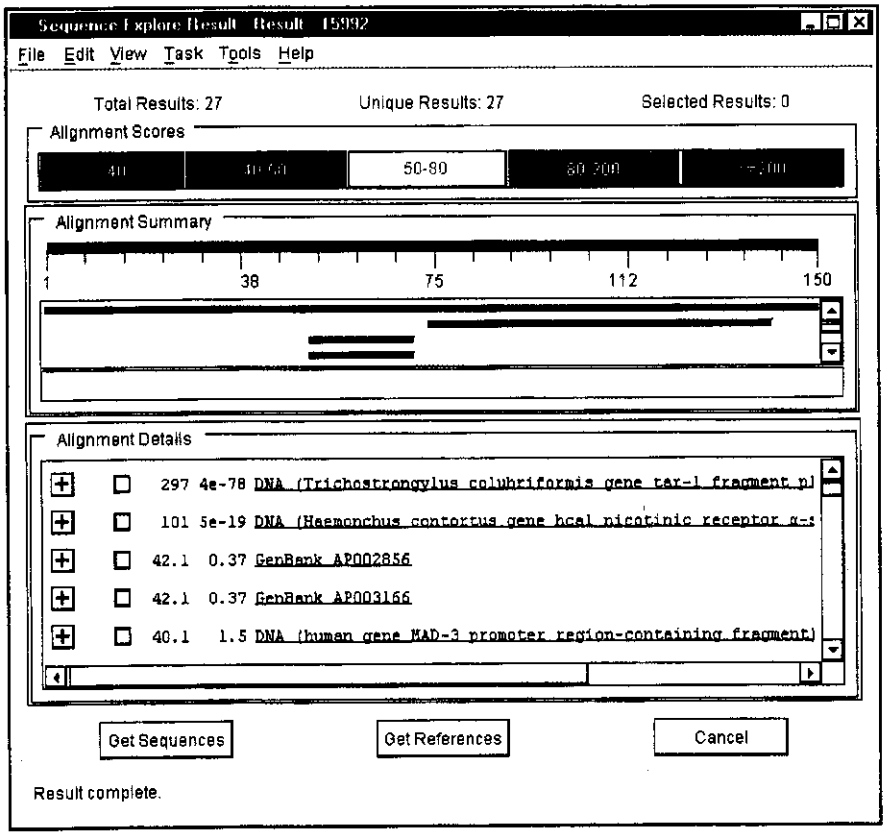



**Figure 4.15** Sample of answer from BLAST search. SciFinder screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

of Spotfire DecisionSite have additional opportunities through the interface. Spotfire *http://www.spotfire.com* is a leading provider of eAnalytic applications and services. SciFinder2001 enables scientists to export physical property data from SciFinder to DecisionSite for further analysis, and selected compounds may be returned to SciFinder for further exploration.

Panorama looks at index terms in records in CAPLUS and MEDLINE and, in particular, enables intersections of index terms to be examined. That is, Panorama enables the user to find records in which two index terms are present and thus develops to a far greater extent the ability to **Analyze References** by **Index Term** considered in Section 2.5.4 of Chapter 2.



However, think about the problem. If Panorama is a tool that helps with the visualization of large answer sets and effectively selects index terms, then large answer sets would have very large number of index terms. The user may well have an even more complex problem sorting through index terms than sorting through the original records! Panorama deals with this problem by first sorting index terms into categories. Help with understanding categories is available through the **Help** menu at the top of the SciFinder screen.

Panorama may be considered as a 'three-dimensional' search aid. In the first dimension is the initial query, while the index terms in the rows and the columns make up the other dimensions. For example, the original query may focus on 'Chinese traditional medicines', the rows may include the substances, and the columns may include any number of index headings such as those for various diseases or treatments. Of course, such searches may always be performed individually through any number of different **Explore by Research Topic** queries. With Panorama, the complex matrix may be determined through a single procedure, thereby giving not only a snapshot of the whole subject but also indicating the number of answers with specific interconnections.

As an example, it is well known that there are links between genetic makeup and cancer, so a question may be 'Can Panorama help with the retrieval of information where cancer is linked to the various human chromosomes?'

First, an answer set needs to be obtained, so the search 'I am interested in ''human chromosome with cancer'' ' is conducted. This gives more than 8000 records in which it is noted that SciFinder automatically searches for many synonyms within the concept 'cancer' (neoplasms, carcinomas, etc.). More than half of these records are from CAPLUS, which again emphasizes the point that workers in this field need to consider CAPLUS as well as MEDLINE.

Panorama ultimately presents a two-dimensional spreadsheet in which one set of index terms is in rows and a second set is in columns. Therefore, once Panorama is clicked, it first looks at all the index terms in the records and presents them in categories (left-hand column in Figure 4.16). The categories are ranked in order of entries, so Figure 4.16 informs that most records are in the category 'Genetics & protein chemistry', and more specifically in the subcategories 'Genetics', 'Nucleic acids', 'Proteins & peptides', and 'Protein & peptide topics'. The user may scroll down through other categories, namely, 'Biology', 'Physical chemistry', and so on.

However, Panorama is a 'big picture' analysis tool and users should try broad options (sets with large reference counts) first. Therefore, once

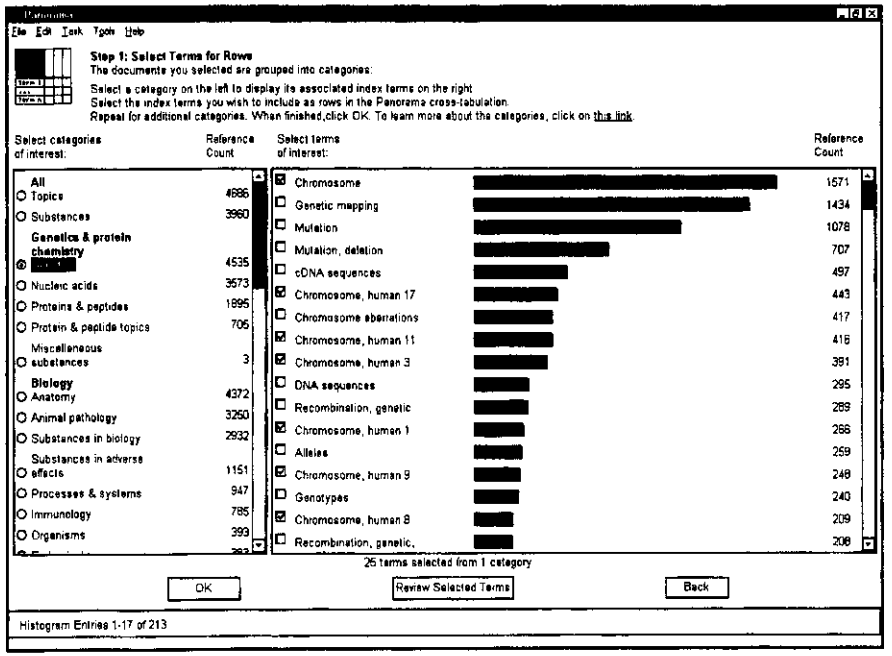



**Figure 4.16** Panorama first requires selection of terms from within a category. These terms will form the rows in the spreadsheet. SciFinder screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

subcategory 'Genetics' is selected, the histogram on the right-hand side of Figure 4.16 is presented, but selection of other subcategories would give different histograms that could be evaluated. Now, it is apparent that there are index terms for each of the human chromosomes, so these are checked in turn. Finally, 'OK' is clicked.

Next, Panorama asks for headings for the columns in the spreadsheet and again, any category may be selected. In Figure 4.17, All:Topics is selected, the histogram is presented, and the user checks those headings of interest. When 'OK' is clicked, an Excel spreadsheet is produced, and it may be manipulated in any way required. (SciFinder does not allow the initial spreadsheets to be edited; the spreadsheet should be copied and pasted into a new Excel document if further manipulations are required). Figure 4.18 shows the result of sorting the spreadsheet into the various chromosomes and indicates, for example, that there are 87 records with Index Terms 'Mammary gland, neoplasm' and 'Human chromosome 17'. If the user clicks on this cell, the 87 records will be displayed.



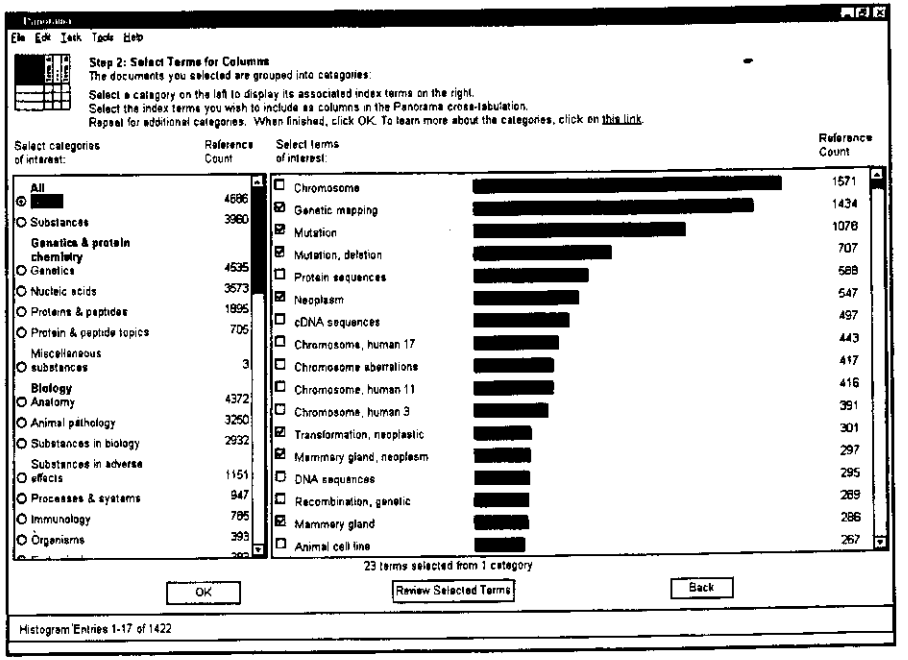



**Figure 4.17** Panorama next requires selection of terms for the columns in the spreadsheet. SciFinder screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

Further manipulation of the data may be undertaken through Excel, and as an example, Figure 4.19 shows a histogram of the hits for the various chromosomes. Remember that the initial answer set was obtained from a search on human chromosome with cancer, so this histogram gives a broad picture of connections between the terms.

If the original answers had been processed with **Analyze References** by **Index Term**, followed by selection of index term 'Chromosome, human 17', then all the original records with this term would have been produced. A further analysis of this subset with **Analyze References** by **Index Term** 'Mammary gland, neoplasm' would give the same 87 records, and the process may be repeated for different intersections of index terms. Alternatively, if a number of boxes found from the first **Analyze References** by **Index Term** were checked, then records retrieved would contain any one of these index terms (effectively the 'or' operator applies). Subsequent analysis with the selection of various terms would result in answers in which any of these terms ('or' operator) was present in the original analysis set ('and' operator).

| Chromosome | Mutation | Genetic mapping | Mammary gland | Neoplasm | Prostate gland | Mutation, deletion | Transformation, neoplastic | Lung, neoplasm | Intestine, neoplasm | Mammary gland, neoplasm | Ovary, neoplasm | Liver, neoplasm | Kidney, neoplasm | Lung, neoplasm, small-cell carcinoma | Stomach, neoplasm | Leukemia | Nerve, neoplasm, neuroblastoma | Mammary gland, carcinoma | Carcinoma | Pancreas, neoplasm | Prostate gland, neoplasm | Intestine, neoplasm, colorectal carcinoma | Liver, neoplasm, hepatoma | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Chromosome | 536 | 494 | 228 | 226 | 155 | 16 | 137 | 135 | 123 | 10 | 86 | 76 | 67 | 4 | 53 | 53 | 6 | 10 | 39 | 37 | 5 | 3 | 8 | 2507 |
| Chromosome, human | 8 | 5 | | 9 | | 6 | 1 | 1 | | 2 | | | 1 | | | | | 3 | 2 | | 1 | | 2 | 41 |
| Chromosome, human X | 13 | 18 | | 10 | | 11 | 4 | 1 | | 5 | 4 | | 6 | 3 | 3 | | 1 | 5 | 2 | | 2 | 1 | 3 | 92 |
| Chromosome, human 1 | 40 | 97 | 3 | 22 | 5 | 76 | 7 | 2 | 2 | 30 | 1 | 1 | 9 | 13 | 3 | | 41 | 17 | 10 | 3 | 11 | 8 | 8 | 409 |
| Chromosome, human 2 | 24 | 34 | | 7 | 1 | 19 | 4 | 3 | 1 | 4 | 2 | 1 | 4 | 7 | 4 | 1 | 5 | 3 | 3 | 1 | 3 | 2 | 8 | 137 |
| Chromosome, human 3 | 75 | 133 | 4 | 31 | | 151 | 16 | 31 | 2 | 15 | 8 | 1 | 4 | 57 | 6 | 2 | | 16 | 8 | 4 | 3 | | 2 | 569 |
| Chromosome, human 4 | 16 | 24 | | 2 | | 31 | 3 | 2 | 1 | 1 | | | | 4 | 4 | | 1 | 6 | 3 | | 4 | | 12 | 114 |
| Chromosome, human 5 | 30 | 30 | | 8 | | 50 | 5 | 6 | 1 | 2 | 2 | | | 11 | 6 | | | 6 | 7 | 1 | 4 | 11 | 6 | 186 |
| Chromosome, human 6 | 19 | 36 | 1 | 6 | 1 | 53 | 8 | 3 | | 13 | 3 | | 1 | 3 | 3 | 1 | | 10 | 4 | 1 | 4 | | 4 | 174 |
| Chromosome, human 7 | 29 | 42 | | 16 | 2 | 41 | 7 | | | 10 | 3 | 1 | 1 | 1 | 3 | 1 | | 7 | 10 | 2 | 12 | 3 | 5 | 196 |
| Chromosome, human 8 | 38 | 85 | 1 | 14 | 6 | 68 | 9 | 4 | | 22 | 2 | 1 | 2 | 7 | 5 | | 1 | 19 | 9 | 3 | 35 | 9 | 15 | 355 |
| Chromosome, human 9 | 59 | 72 | 2 | 17 | 3 | 114 | 13 | 9 | | 7 | 6 | | 2 | 7 | 3 | 3 | 2 | 9 | 3 | 4 | 4 | 1 | 7 | 347 |
| Chromosome, human 10 | 35 | 63 | 1 | 8 | 2 | 45 | 3 | 5 | | 4 | 1 | | | 9 | 2 | 1 | | 6 | 4 | | 15 | 1 | 6 | 211 |
| Chromosome, human 11 | 62 | 135 | 7 | 36 | 3 | 109 | 30 | 12 | 2 | 43 | 14 | 2 | 3 | 6 | 7 | 3 | 4 | 39 | 17 | 5 | 13 | 4 | 9 | 565 |
| Chromosome, human 12 | 14 | 41 | 1 | 14 | 1 | 15 | 3 | 1 | | 4 | 1 | | | 2 | 3 | | 2 | 6 | 4 | 2 | 5 | | 2 | 121 |
| Chromosome, human 13 | 38 | 35 | 1 | 12 | 1 | 57 | 9 | 6 | 1 | 27 | 5 | | | 12 | 4 | 1 | 1 | 8 | 6 | 2 | 10 | 2 | 7 | 245 |
| Chromosome, human 14 | 10 | 26 | | 6 | | 20 | 4 | 2 | | 3 | 4 | | 1 | 3 | 3 | 1 | 2 | 1 | 1 | | 1 | | 2 | 89 |
| Chromosome, human 15 | 7 | 16 | | 2 | 1 | 16 | 2 | 3 | | 1 | 1 | | | 2 | 4 | | 2 | 2 | 2 | 1 | | | | 63 |
| Chromosome, human 16 | 29 | 42 | 1 | 5 | 1 | 42 | 5 | 3 | | 20 | 3 | 1 | 1 | 5 | 2 | | 14 | 21 | 2 | | 8 | 1 | 12 | 204 |
| Chromosome, human 17 | 124 | 125 | 2 | 20 | 2 | 141 | 10 | 11 | 2 | 87 | 35 | 1 | 2 | 14 | 10 | 2 | 6 | 39 | 15 | 4 | 14 | 23 | 16 | 705 |
| Chromosome, human 18 | 40 | 27 | | 3 | | 53 | 1 | 3 | 1 | 9 | 1 | | | 2 | 8 | 1 | 3 | 7 | 7 | 5 | 7 | 20 | | 198 |
| Chromosome, human 19 | 11 | 34 | | 3 | | 17 | 2 | 1 | 1 | | 3 | | 1 | 2 | 2 | | | 5 | 2 | 3 | 3 | | | 90 |
| Chromosome, human 20 | 12 | 19 | | 3 | | 25 | 3 | 1 | | 15 | 4 | | | 2 | 4 | 1 | | 9 | 3 | 2 | 4 | 1 | 1 | 109 |
| Chromosome, human 21 | 5 | 15 | | 3 | | 8 | 3 | | | 2 | | | 1 | 1 | 3 | | | 1 | 2 | | 1 | | | 45 |
| Chromosome, human 22 | 16 | 31 | | 4 | | 20 | 2 | 1 | | 1 | 1 | | 1 | 4 | 2 | | | 2 | 3 | | 2 | 1 | 1 | 92 |

**Figure 4.18** Initial Panorama spreadsheet may be manipulated with Excel. Numbers indicate number of records with both of the index terms.



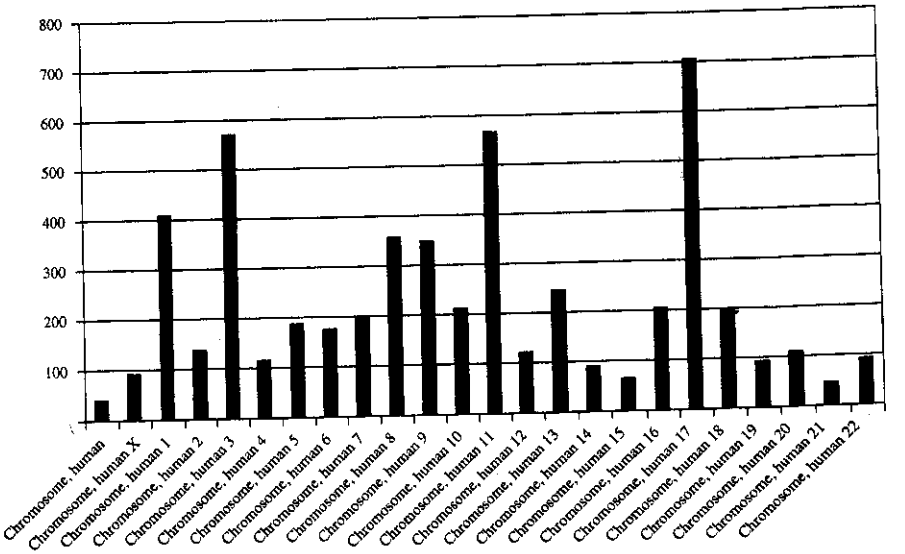




**Figure 4.19** Example of graphical representation possible through Excel. The *y*-axis indicates number of records from the original search that have the chromosome as an index term. It is apparent that many studies relate cancer to human chromosome 17.

Different selections of index terms for rows and columns may have been made from Figures 4.16 and 4.17. For example, selection of terms for human chromosomes from 'Genetics & protein chemistry–Genetics', and then anatomical terms from 'Biology–Anatomy' would give a spreadsheet that linked chromosomes with parts of the anatomy. Since all these answers ultimately came from a search that required a term related to cancer, information on chromosomes or anatomy with cancer would result.

The additional power of Panorama is thus evident!

## 4.9 Current Awareness

Keeping abreast with published literature may be achieved in different ways. One way is to browse the table of contents of serials through full text document delivery services or through Tables of Contents that are now provided by publishers directly through their Web sites. Another way is to manually update searches on a regular basis or to have updates delivered automatically after a certain time. Scientists may employ all these methods through SciFinder, although the last option, Keep Me Posted, currently is not available to SciFinder Scholar users.

### 4.9.1 Browse Table of Contents

SciFinder Scholar offers a Table of Contents option, and presently, more than 1660 journals are covered. It is a simple matter to click **Browse Table of Contents** from the main screen (Figure 2.1), and the user is presented with the list of titles (Figure 4.20).

Although the list may be scrolled down, it is more convenient to go to **Edit**, then **Find** from the pull-down menu, and enter one or two words in the title of the journal of interest. SciFinder Scholar shows the first match for the word(s) used, but if this is not the correct journal, then the user goes to **Edit**, then **Find again** (or **Ctrl G**), and the next match is displayed. After a few **Ctrl G**s, the journal is found quickly. For example, to find *The Journal of the American Chemical Society* (the actual title for this journal in Browse Table of Contents through SciFinder Scholar), the user goes to **Edit**, then **Find**, and enters 'american chemical'. SciFinder Scholar first takes the user to the Book of Abstracts of the American Chemical Society, but **Ctrl G** then takes the user down to the journal required (Figure 4.21)

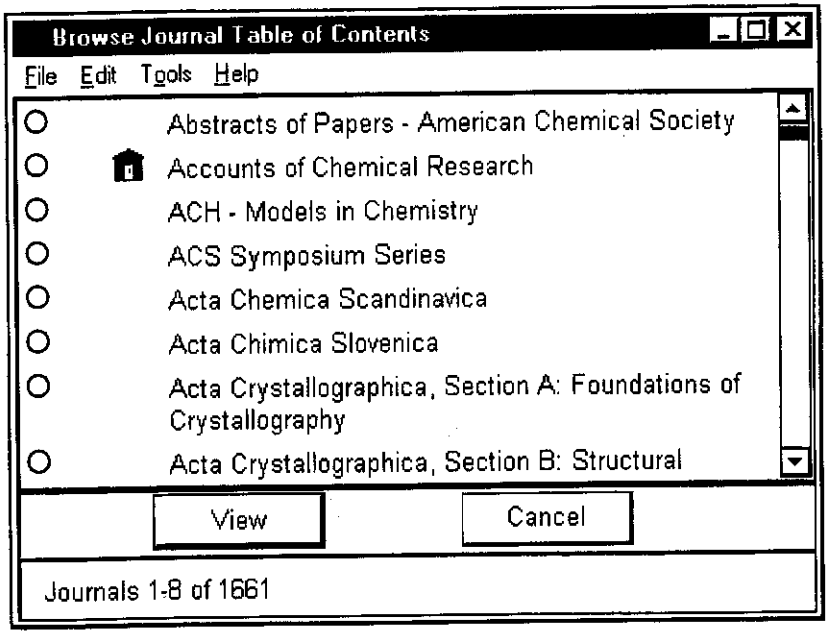



**Figure 4.20** **Browse Table of Contents** gives journal lists. **Edit** then **Find** provides a quick way to scroll to the required journal. SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.



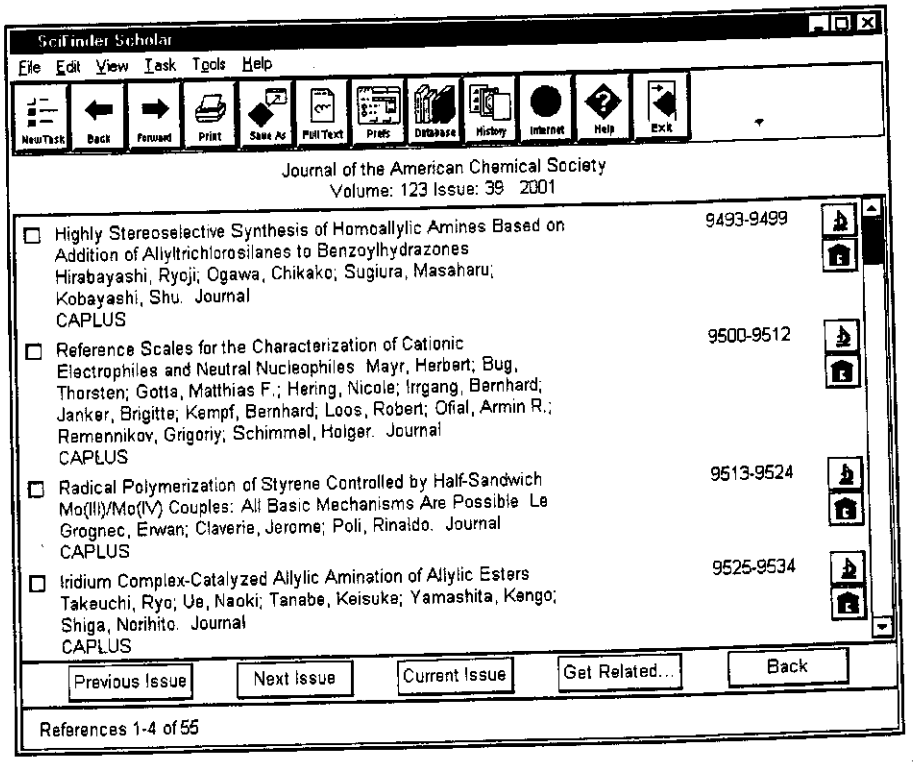



**Figure 4.21** Example of Table of Contents from the Journal of the American Chemical Society. SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

From this point, the user may click the check box, the microscope, the house or computer icons, or the Previous Issue, Current Issue, Next Issue, or Back buttons as required. Depending on the issue being viewed, the user may be able to additionally look at the full record (including citations) in CAPLUS.

### 4.9.2 Keep Me Posted

SciFinder users may request automatic updates of searches either through the initial SciFinder screen or through the icon at the bottom right in most screens. The process is entirely intuitive and the user is prompted for the various pieces of information required.

The main issue to be aware of is that a retrospective search strategy that may cover records for more than 30 years may not necessarily be the best strategy to use for update searches that may cover records for only a

couple of weeks. Thus, an answer set of 3000 records in the full database may need substantial refining in order to achieve a manageable answer set. However, if these answers are accumulated over a 30-year period, this amounts to an average of around two records per week. Accordingly, SciFinder users may need to develop more general strategies for Keep Me Posted profiles.

## 4.10 Exercises

4.1 Find all records for articles written by:

4.1.1 Eric de Clercq (Rega Institute);

4.1.2 Roeland J M Nolte (University Nijmegen);

4.1.3 Andre Lubino (CRNS) (that's how his name sounds, anyway!);

4.1.4 Ian F C McKenzie (Austin Research Centre).

4.2 Find the record for the patent WO99/53915. Has this patent been cited recently?

4.3 Find records that report publications from:

4.3.1 3M Japan;

4.3.2 The Commonwealth Scientific and Industrial Research Organization (CSIRO) in Australia;

4.3.3 The Chemical Abstracts Service;

4.3.4 Cambridge University, UK.

4.4 Find recent citations to publications by:

4.4.1 Engelbert Zass;

4.4.2 Karl M Kadish.

What are the parameters (restrictions) under which these searches were performed?

4.5 What esters (not lactones) are mentioned in records for the famous spirit, whisky? What analytical studies (in whisky) have been reported for these substances?

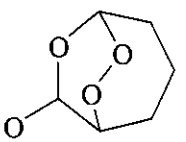
4.6 Find substances that have been reported in patents on photodynamic therapy. How many of these substances contain rings with four carbons and one nitrogen atom? Which companies have published patents on the use of these substances in photodynamic therapy? Are any of these substances commercially available? What ring skeletons do these commercially available substances have?



4.7 Locate the records in CAPLUS and MEDLINE for the original article on BLAST titled 'Basic local alignment search tool' by Stephen Altschul *et al. Journal of Molecular Biology*, **215**, 1990, 403. Explore further literature on the description and application of the algorithms by looking through some of the records for articles that cite the original article (**Get Related: Citing References**).

4.8 Find records that report Lipinski Rules. What publications have cited these records?

4.9 An important part of the structure of the antimalarial artemisinin is the bicyclic structure (A). What compounds containing this structure conform to 'Lipinski Rules' (<6 hydrogen-bond donors, <11 hydrogen-bond acceptors, molecular weight <500 and calculated log $P$ < 5)? How many of these substances are at least soluble in aqueous solution (pH range not important)?




(A)

4.10 Tracing the source of drugs by their 'profiles' is an important forensic issue. Find records in which the concepts 'profiling' and 'heroin' are closely associated. What synonyms have been searched? What substances are reported in these records and how many have the morphine skeleton? Find analytical methods that use mass spectrometric techniques for these substances.

Chapter 5

# Additional Search Strategies

## 5.1 Introduction

The ability to search for entries in titles, abstracts, and indexing provides the user with multiple access points to information, while the ability to search in different databases helps increase the breadth of coverage. However, as opportunities increase, so also does the need for the scientist to make informed choices. This chapter discusses some of the alternative ways to approach problems and further elaborates on the importance of SciFinder Scholar as a research tool.

## 5.2 Further Issues with Finding Information on Substances

### *5.2.1 Research Topic Versus Substance Identifier*

It is apparent from Chapters 2 and 3 that information on substances may be found either by entering the substance along with the additional topics in **Explore by Research Topic** or by starting with **Explore by Chemical Substance**, then proceeding through a variety of steps to the references. That is, if the user is interested in information on the antibiotic/immunosuppressant cyclosporin A and cancer, alternative approaches may be:

(a) **Explore by Research Topic**
'I am interested in ''cyclosporin A with cancer'' ' or
(b) **Explore by Chemical Substance**, then **Substance Identifier**
'cyclosporin A', then **Get References**, followed by **Refine References** with 'cancer'.

The preferred option depends very much on the actual search, and generally, both options should be investigated. As is often the case, it is best to do the experiment! In this case, when option (a) is chosen, SciFinder Scholar indicates that there are just over 500 records in which the concepts **cyclosporin** and **cancer** are 'closely associated' and just over 2000 records in which the concepts are 'anywhere in the reference'. A quick check on some of the answers indicates that the usual synonyms for cancer (neoplasm, carcinoma, etc.) are searched, while the terms retrieved for cyclosporin included the Chemical Abstracts Service (CAS) Registry Number for cyclosporin A (59865-13-3) and the MEDLINE index entry cyclosporine. (As mentioned in Chapter 2, it is important to check that possible index terms have been searched, and this example indicates how well SciFinder Scholar generally performs). Table 5.1 summarizes the outcome of the search and some of the issues involved.

When option (b) is chosen, the substance cyclosporin A is easily found and **Get References** affords more than 23 000 records. When these are refined with 'cancer', there are almost 1400 answers of which slightly more are in MEDLINE than in CAPLUS. Table 5.2 summarizes the issues, and in particular, it should be noted that a primary difference is that only the CAS Registry Number for cyclosporin A is searched through this process.

A further option now worth considering is an **Explore by Research Topic** 'I am interested in ''59865-13-3 with cancer'' ' since this will give a candidate answer set in which the two terms are 'closely associated'. In fact, there are almost 200 records for this candidate and of course, all are from CAPLUS (since CAS Registry Numbers in MEDLINE are in a separate field/sentence).

Therefore, perhaps an **Explore by Research Topic** 'I am interested in ''cyclosporin A (59865-13-3) with cancer'' ', and then selection of the 2000 and more records in which the different concepts are 'anywhere in the reference' may be a useful starting point. From there, the powerful analyze and refine options may be used to select more specific answers.

While this example outlines some searches for the topic 'cyclosporin A with cancer', searches under options (a) or (b) for different substances or topics will naturally turn up other variations. The key issues for the searcher are the impacts of specifically searching for CAS Registry Numbers, of how the algorithm searches under **Explore by Research Topic** for the various 'words' in the name for the substance (e.g. **Explore by Research Topic** 'I am interested in ''4-nitropyridine. . . .'' ' will identify concepts '4' and 'nitropyridine', and even candidates where they are 'closely associated' may contain the concepts in very different parts of the sentence), of how those 'words' appear in actual records (e.g. searches on the words 'benzoic acid'



**Table 5.1** Summary of issues relating to searching for information on substances through **Explore by Research Topic**. Example: I am interested in 'cyclosporin A with cancer'

| | Search issue | Example |
|---|---|---|
| 1 | When a substance is entered, the term is searched as a word, and the full algorithm is implemented | A search on 'cyclosporin A' ignores the 'A' since it is a word that cannot be searched. The concept searched is 'cyclosporin' and the algorithm applies truncation. The index term in MEDLINE 'cyclosporine' thus is a hit term |
| 2 | The CAS Registry Number will be searched only if SciFinder Scholar finds an exact match of the term entered with a complete name in the substance database | 'Cyclosporine' happens to be a synonym for cyclosporin A in REGISTRY, so in this case, the CAS Registry Number 59865-13-3 is included within the concept. However, the inclusion of the CAS Registry Number should always be checked |
| 3 | Candidates with the concepts 'closely associated' are displayed. This affords a level of precision, although the choice should be carefully made (e.g. MEDLINE index terms are never 'closely associated' with words in titles or abstracts or with other index terms in the record) | A search on 'cyclosporin A with cancer' gives 502 records with the concepts 'closely associated' (348 and 154 in CAPLUS and MEDLINE, respectively) |
| 4 | Candidates with the concepts 'anywhere in the reference' are displayed | A search on 'cyclosporin A with cancer' gives 2180 records with the concepts 'anywhere' (915 and 1265 in CAPLUS and MEDLINE, respectively) |
| 5 | Searching for the word may retrieve records for other substances | An outcome of issue 1 above is that any record that includes the words 'cyclosporin B', 'cyclosporin C' (in fact, all cyclosporins A to Z are known!) will be retrieved. Such an outcome may thus help achieve comprehension, but precision may be reduced |
| 6 | Searching for the word may retrieve unwanted hits through application of truncation | In this case, truncation to 'cyclo. . .' did not occur (!), but in other cases, truncation may cause difficulties (see Section 2.3.2 of Chapter 2) |

will retrieve records with entries 'benzoic acid esters', etc.), and of the need to search for concepts 'closely associated'.

In summary, searching for substances *by name* in **Explore by Research Topic** may lack precision but may aid comprehension. On the other hand, finding substances by name, starting from **Explore by Chemical Substance**, will always employ CAS Registry Numbers in subsequent steps. These are precise search terms that also aid in comprehension since they may

**Table 5.2** Summary of issues relating to searching for information on substances by name through **Explore by Chemical Substance**. Example: cyclosporin A

| | Search issue | Example |
|---|---|---|
| 1 | **Explore by Chemical Substance**: **Substance Identifier** searches the exact name in the substance database (Section 3.6 of Chapter 3) | Cyclosporin A (59865-13-3) is easily retrieved by any one of a number of simple names (Cyclosporin A, Cipol N, Cyclosporine, Sandimmun, etc.) |
| 2 | Answers are substances, and further steps need to be followed to get references | More than 23 000 records are retrieved when the CAS Registry Number 59865-13-3 is searched in CAPLUS and MEDLINE |
| 3 | **Get References** retrieves all references or references in which a number of concepts (Figure 3.21) are closely associated with the CAS Registry Number. The latter allows precision searches in CAPLUS; there are different applications in MEDLINE (Section 5.2.2) | The crossover from REGISTRY to either CAPLUS or MEDLINE uses CAS Registry Numbers only as search terms. Restricting references by concepts in crossover uses CAS Roles or text-modifying phrase terms as search terms |
| 4 | Bibliographic answers are refined by Research Topic at the 'anywhere in the reference' level, and 'closely associated' options are not possible except for those included in issue 3 mentioned above | It is not possible to start with **Explore by Chemical Substance** 'cyclosporin A' and subsequently 'closely associate' answers with 'cancer'. However, choosing the more than 23 000 answers mentioned above and then **Refine References** with **Research Topic** 'cancer' gives almost 1400 records (630 CAPLUS, 760 MEDLINE) |

effectively cover many names for a substance and may be the only entry point for a substance in a record (e.g. the CAS Registry Number 158012-18-1 in Figure 1.3).

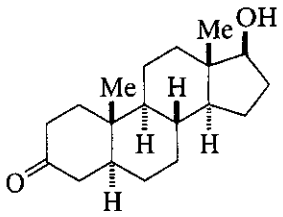
### *5.2.2 CAS Registry Number Crossover*

When 'all references' (Figure 3.21) is chosen in the file crossover, records with the CAS Registry Numbers in both CAPLUS and MEDLINE are retrieved. On the other hand, if options under 'only of the following types' are chosen, records will be restricted. Precisely what records are retrieved depends on the actual search, but the basic principle is that answers are retrieved through searches of CAS Registry Numbers linked with appropriate CAS Roles, text-modifying phrases, MEDLINE Index Headings, or allowed qualifiers (AQ).

Since indexing policies in the two bibliographic databases differ, SciFinder Scholar needs to interpret the crossover in different ways. For example, a



name search under **Explore by Chemical Substance** for dihydrotestosterone retrieves the anabolic steroid (**1**). When all references for this substance are requested, around 13 500 bibliographic records are retrieved (around 7400 from CAPLUS and 6100 from MEDLINE). Approximately 3500 records are duplicates.




**(1)**

While the index entry in CAPLUS for dihydrotestosterone is the CAS Registry Number 521-18-6, the index entry in MEDLINE is STANOLONE, and the CAS Roles in CAPLUS are grouped differently from the Allowed Qualifiers for the substance in MEDLINE. The CAS Role ADV (Adverse Effect, including toxicity) is applied to the CAS Registry Number when information 'only of the following types: **Adverse effect, including toxicity**' is used to retrieve answers in CAPLUS. However, SciFinder Scholar needs to consider for the same search a number of MEDLINE headings including 'drug effects' and 'adverse effects,' and the allowed qualifiers TO (toxicity) and AE (adverse effects).

When 'only of the following type' **Preparation** is chosen, approximately 1200 records are retrieved and more than 1000 of these are from MEDLINE. Surely, there cannot be more records for the synthetic preparation of dihydrotestosterone in MEDLINE than in CAPLUS? Of course there are not, but the example illustrates the point that the way the databases are indexed and the way SciFinder Scholar performs this crossover from REGISTRY are different.

Table 5.3 gives some examples of the number of records retrieved when the substance dihydrotestosterone (CAS Registry Number 521-18-6) is found through **Explore by Chemical Substance** and when various types of information 'only of the following types' are chosen.

Finally, as seen in Figures 1.3 and 1.5, the indexing of CAS Registry Numbers in the databases also differs in a fundamental way. While the policy of CAS indexers is to enter CAS Registry Numbers for the precise substance, MEDLINE indexers may use a single CAS Registry Number to cover some closely related substances. For example, the CAS Registry Number 521-18-6

**Table 5.3** Number of records retrieved when refinement options are chosen within **Get References** for dihydrotestosterone

| Option chosen: Get References, then... | Number of records | |
|---|---|---|
| | CAPLUS | MEDLINE |
| All references | 7344 | 6106 |
| Adverse Effect, including Toxicity | 60 | 140 |
| Analytical study | 319 | 2596 |
| Crystal structure | 12 | 0 |
| Preparation | 76 | 1080 |
| Reactant | 288 | 0 |
| Spectral properties | 81 | 0 |
| Uses | 101 | 1383 |

is actually used in MEDLINE for both the 5$\alpha$- and 5$\beta$-dihydrotestosterones, although they are quite different substances.

Accordingly, it helps if careful interpretation of answers is undertaken. As always, a balance between precision and recall needs to be established, and it helps if users try alternatives to ensure that the most appropriate balance is achieved.

## 5.3 Opportunities for MEDLINE Searchers

The MEDLINE database may be searched through many sources. Some options are through PUBMED® *http://www.ncbi.nlm.nih.gov/PubMed* or PUBMED Central *http://www.pubmedcentral.nih.gov*, through some online networks (e.g. Dialog and STN), or local networks like OVID® *http://www.ovid.com* or WebSPIRS® *http://www.silverplatter.com*. The ability to search MEDLINE through SciFinder Scholar offers another interface, which the medical researcher now needs to evaluate. While individual users need to make these evaluations for their own research, some general comments on the implementation through SciFinder Scholar follow.

### *5.3.1 Complementarity of MEDLINE and CAPLUS*

Users of SciFinder Scholar may search both MEDLINE and CAPLUS simultaneously and thus take full advantage of their complementarity.

#### Unique Records

Only by following a most detailed analysis of records in the various medical areas, may a reasonable understanding of the unique coverage in each database



be achieved, but in general, MEDLINE offers better coverage in the areas of clinical, social, and epidemiological medicine. Both databases cover preclinical medicine, but as the studies become more molecular, CAPLUS increasingly becomes the more important resource. Indeed, the Biological Section Codes of CAPLUS contain more than 5.5 million records covering preclinical research including biochemistry, pharmacology, and molecular biology, and more than 280 000 records from the patent literature. The number of patents in these sections is growing rapidly, and more than 53 000 of these patent records were added in the years 1998 to 2000.

#### Unique Indexing

Second, where there is overlap in coverage, the indexing is quite different (e.g. Figures 1.3 and 1.5), and the ability to search two sets of index terms may substantially increase recall.

The importance of indexing cannot be overemphasized. Use of indexing assists with recall since a single index term may cover a variety of terms used by authors and assists with precision since index terms are entered only when they relate to key parts of the original article. For example, Figure 5.1 shows the most commonly posted index headings in the records retrieved in MEDLINE for **Explore by Research Topic** 'I am interested in ''inhibition of HIV replication in humans'' ', and checking key headings of interest will lead to an answer set refined using systematic terminology.

Additionally, identification of index terms may often alert the user to terms that may be added to the initial question. For example, Figure 5.1 indicates headings 'antiviral agents', 'viral replication (drug effects)', and 'HIV-1 (drug effects)', and terms based on these may well be a better way to search for the topic in MEDLINE.

Of course, commonly posted headings are not always of greatest interest, and a review of further headings may yield more specific answers. It helps to remember that index headings are file-specific but there are no problems eventually going through one database once the nature of records in both databases has been examined. It is far better to evaluate answers from both databases rather than to restrict answers to one database initially since the latter process may well cause key answers to be missed.

#### Display and Duplicate Records

By default, **Explore by Research Topic** searches terms in the title, abstract, and index fields of both bibliographic databases. Answers from CAPLUS are presented first, followed by the MEDLINE records, although this order may

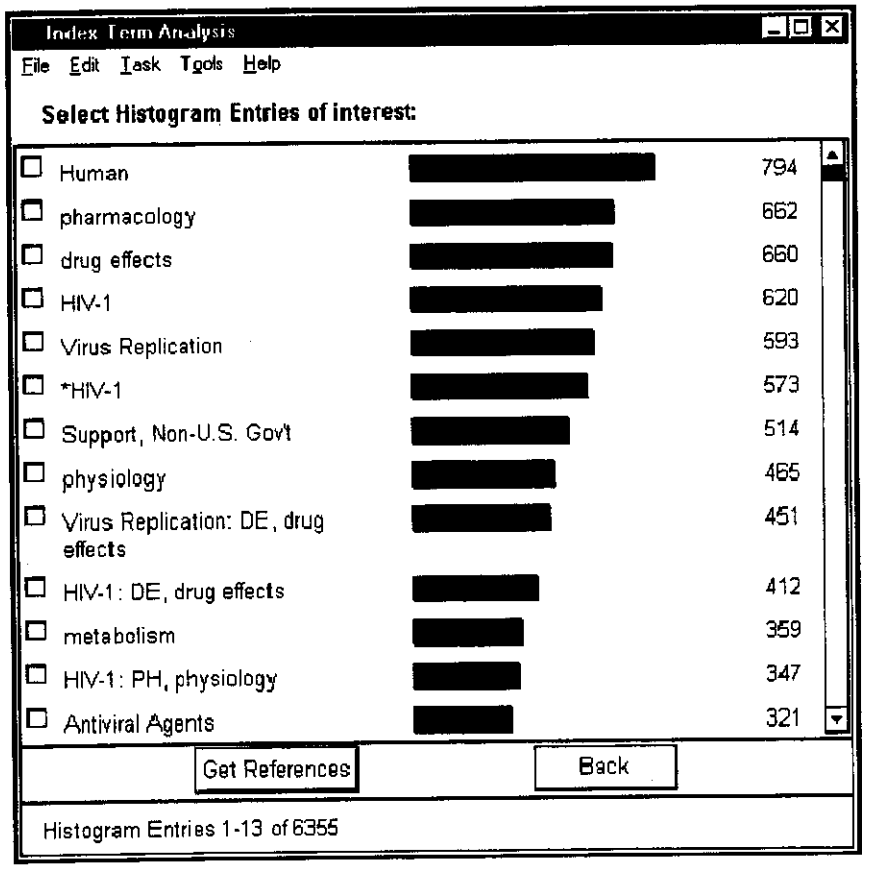



**Figure 5.1** Major index headings for search on 'inhibition of HIV replication in humans' in MEDLINE database. (See Figure 2.11 for index headings in CAPLUS.) SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

be changed. For example, the pull-down tab **View** has options to view in reverse order or title order, and the latter is useful for duplicate removal since it places records for the same original article from the different databases next to each other.

If answers from only one database are required, **Analyze References** by **Database** gives the necessary choices. Currently, it is possible to remove duplicate records by manual selection of records, but in any case, this should be done only after all analyze and refine options, particularly those relating to indexing, have been explored.



### Size of Answer Sets

Since two of the world's largest bibliographic databases are searched through **Explore by Research Topic**, at times, large answer sets may be retrieved initially, and it is important for users to refine answers in a systematic manner. For example, while **Explore by Research Topic** for 'inhibition of HIV replication in humans' gives more than 1200 records in CAPLUS, there are more than 850 records in MEDLINE. Therefore, when both databases are searched, there are more than 2200 records even with the concepts 'closely associated' (Figure 2.3).

Although further concepts may be added in a revised **Explore by Research Topic** query to obtain fewer records, usually it is preferable to work through **Analyze References**. In particular, **Analyze References** by **Index Term** offers a powerful option that takes advantage of the systematic terms added by indexers of each of the databases.

### CA Lexicon® and MeSH® Thesaurus

It is apparent from the above discussion that identification of index terms is only possible at the secondary level in SciFinder Scholar. That is, a bibliographic answer set is obtained first, and the index headings that are identified may be used either directly to narrow the search or as a guide to terms to use in a revised initial search. In fact, this approach serves the end user very well, as it allows the searcher to take advantage of indexing without having to know too many details of index policies in advance.

However, both CAPLUS and MEDLINE have index headings arranged in hierarchical structures (CA Lexicon, Section 1.5 of Chapter 1, and MeSH thesaurus, Section 1.6 of Chapter 1), and examples are given in Figures 1.4 and 1.6, respectively. The advantage of these hierarchical index structures is that searches at broader, narrower, or associated index levels may be conducted, and these may facilitate the retrieval of comprehensive and precise answer sets. Accordingly, it helps if SciFinder Scholar users keep in mind that there are a variety of information retrieval strategies, and some of these are even available through alternative end-user interfaces. Nevertheless, the use of thesaurus capabilities really is a very specialized topic and should be explored with help from information professionals.

## *5.3.2 Complementarity of REGISTRY, MEDLINE, and CAPLUS*

Currently, REGISTRY has more than 1.6 million protein sequences and more than 21.6 million nucleic acid sequences, including more than 18.9 million substances from GENBANK® *http://www.cas.org/ONLINE/DBSS*

*/genbankss.html*. There also are more than 56 000 different CAS Registry Numbers in MEDLINE, and they appear in more than 4 million (almost one-third!) MEDLINE records. Meanwhile, CAPLUS has more than 5 million records that list the CAS Role: Biological Study, and the numbers of records with the specific Roles: Adverse Effect and Therapeutic Use are more than 300 000 and 400 000, respectively. This data gives an idea of some relationships between the three major databases in SciFinder Scholar.

Substances may be retrieved in SciFinder Scholar through name, formula, or structure searches, and the latter in particular, presents opportunities that are not readily available to MEDLINE searchers. The exception is that structure searches may be performed through a variety of files on the STN network, and

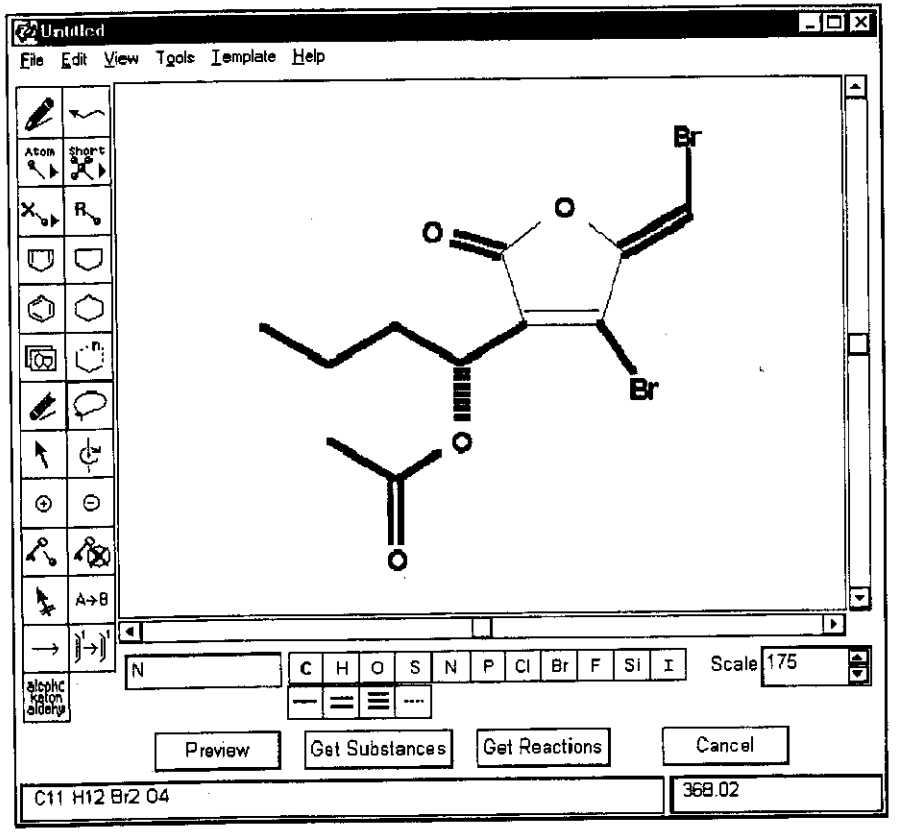



**Figure 5.2** Structure of furanone known to have activity against gram-positive bacteria. SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.



the CAS Registry Numbers in the answer set may subsequently be searched in the MEDLINE file. As an example of the importance of structure searches in the medical sciences, consider the situation of a scientist who is interested in the use of furanones as inhibitors for gram-positive bacteria. This interest may have been developed either through prior knowledge of the literature or perhaps through an **Explore by Research Topic** 'I am interested in ''the inhibition of gram-positive bacteria'' ' that produces, *inter alia*, a patent (WO99/53915 titled 'Inhibition of gram-positive bacteria' by Kjelleberg *et al.*) that describes such a use for the furanone with CAS Registry Number 63025-21-8. The question is 'What other furanones have activity against gram-positive bacteria?'

When this Registry Number is copied and pasted into the structure drawing program (Section 3.4.1 of Chapter 3), the structure in Figure 5.2 appears. This may then be modified to give a more general furanone structure (Figure 5.3), which in this case has the ring locked only.

When **Get Substances** followed by 'a substructure of a more complex structure' is chosen, more than 5000 substances are retrieved. However, users should not necessarily be put off by large answer sets, particularly in SciFinder Scholar, which has excellent ways to analyze/refine answers. In this case, while the substances may be narrowed with the help of the various analysis tools discussed in Section 3.5 of Chapter 3, the ultimate intention is to restrict *references* to those that mention gram-positive bacteria, and at this final stage, probably relatively few references will result. Therefore, **Get References** is chosen and this gives more than 2500 references, all of which have at least one of the CAS Registry Numbers in the substructure answer set. **Refine References** with the **Research Topic** 'gram positive' gives an answer set of 30 records of which 24 are from CAPLUS and six from MEDLINE. Part of one of the records is shown in Figure 5.4.

Even though this is an excellent result, it helps if the user considers the possible limitations of the search and considers what alternative strategies may have been used. In this case, a precise search on a part structure was undertaken and this would have comprehensively and precisely covered the 'furanone' concept in the original query. Of course, the answers would be restricted to those records that listed the CAS Registry Numbers, but it is reasonable to assume that if something new about the antibiotic property of furanone derivatives was mentioned in the original article, then the CAS Registry Numbers would have been indexed.

However, the refinement 'gram positive' may not have been comprehensive and, in particular, refinement with the names of specific bacteria may be considered. Once again, clues on how to proceed may be gained through

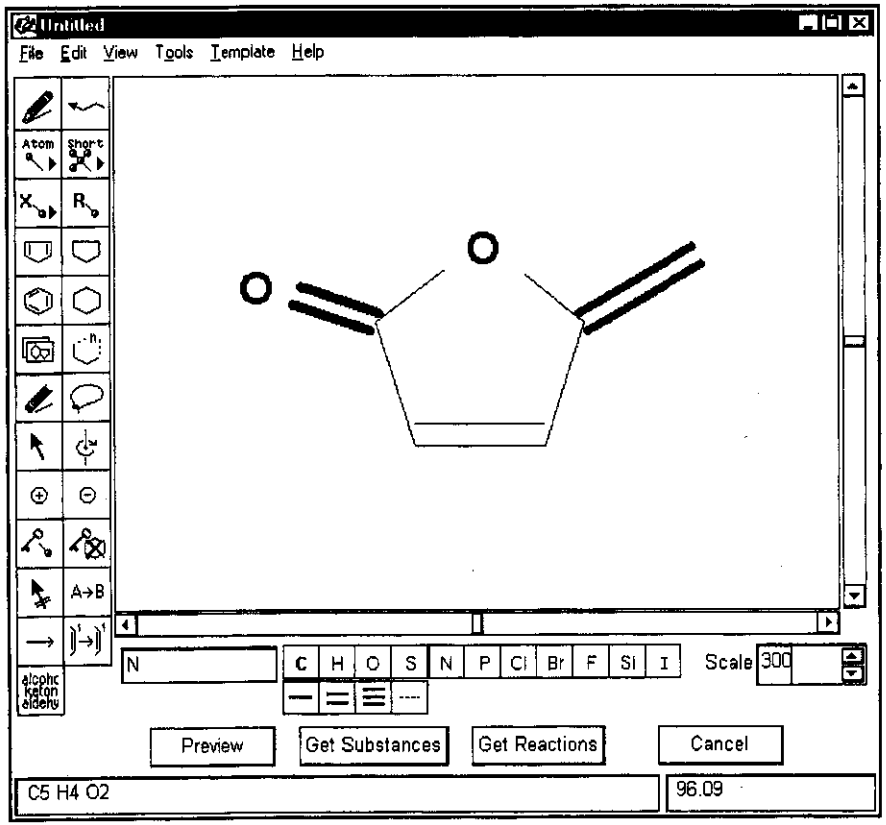



**Figure 5.3** Modified furanone structure for substructure search. SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

SciFinder Scholar and, in particular, through analysis of the 30 answers by Index Term. When this is done, the names of a number of specific bacteria become apparent and these may be entered as an alternative refinement for the initial answer set. Indeed, the refinement of these records with terms 'Enterobacter or Staphylococcus or Streptococcus' yields 13 new records. The final result thus is a set of 43 references. This is an even better result!

This example illustrates the value of the ability to search for substances in REGISTRY in a *precise* manner, and then to search for information on these substances in both CAPLUS and MEDLINE. Effectively, this links the CAS substance and bibliographic databases with the MEDLINE database in a unique way, so medical scientists now have opportunities to find more



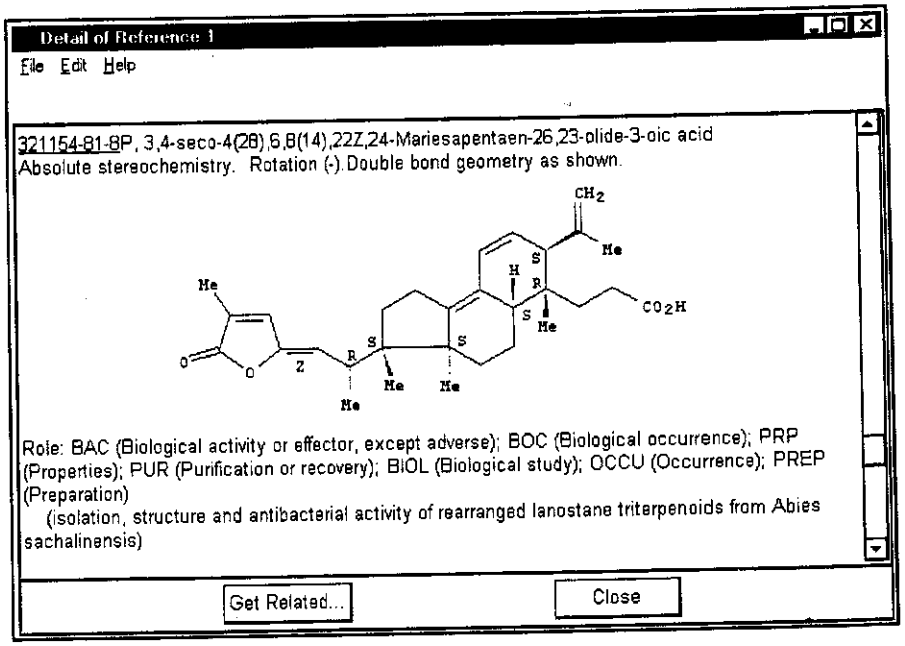



**Figure 5.4** Example of bibliographic answer for search for furanones with activity against gram-positive bacteria. Copyright the American Chemical Society and reproduced with permission.

precise and comprehensive information and creative ways to obtain leads to new areas.

### 5.3.3 *The Search Interface and Search Opportunities*

Through the intelligent interface described in Chapter 2, SciFinder Scholar guides the user and considerably assists with the retrieval of relevant records. This intelligent interface also helps with the retrieval of records for authors (Section 4.2 of Chapter 4) and companies (Section 4.4 of Chapter 4). There is no doubt that once the overall philosophy and function of SciFinder Scholar is understood, scientists may very quickly and effectively accomplish excellent search results.

Further, the manner in which SciFinder Scholar integrates the databases provides many search opportunities, particularly through the ability to **Get Related** answers. This is important not only with citation searching but also with the ability to get related substances. Thus, if the medicinal scientist is interested in substances to which the malaria parasite is not resistant, then **Explore by Research Topic** 'I am interested in "malaria with resistant"'

affords around 1000 records in CAPLUS. These records may be analyzed by CAS Registry Number to obtain a histogram of substances or the substances may be retrieved through **Get Related** and then **Substances**. This gives a REGISTRY answer set that in turn may be analyzed or refined.

In summary, SciFinder Scholar offers many opportunities for those who traditionally have used MEDLINE!

## 5.4 Searching for Substances in the Biological Sciences

Although searching for organic and inorganic substances presents a number of challenges, very often the scientist is dealing with a discrete substance, and some of the options discussed in Chapter 3 will usually produce excellent answers. However, searching for 'biological' substances often presents greater challenges because the exact nature of the substance may be unknown, or if it is known, it may be difficult to describe. For example, how are genes, plasmids, or immunoglobulins described? How may slight genetic modifications be retrieved? Transport proteins? Antibodies? Receptors? Complementary DNA?

It *is* possible to search for individual 'biological' substances directly in REGISTRY using the techniques in Chapter 3, and the scientist should first explore some alternatives. However, if more general searches are required, different techniques may need to be employed. For some general information on searching biosequences, see *http://www.cas.org/SCIFINDER/SCHOLAR/SOLUTIONS/bioseqover.html*.

When contemplating how to achieve this, it helps if the scientist goes back to basic principles, that is, the entries in bibliographic databases are written in part by authors and in part by indexers. The challenge is to anticipate terms used by authors or to know how the field is covered by indexers. Actually, neither option is that daunting since scientists are aware of the author terminologies in their area and since indexing terms may be identified through **Analyze References** once an initial answer set has been obtained.

Nevertheless, a general understanding of indexing may be helpful, and Table 5.4 summarizes indexing in the field of molecular biology in CAPLUS.[1] The first two columns present *some* general index headings in the field and the approximate time for which they were valid, the third column indicates the

[1] The discussion here relates to CAPLUS partly because this database has broader coverage (e.g. almost one-quarter of the records with index heading 'cDNA sequences' are from patents) and hence should probably be explored in the first instance. However, indexing in MEDLINE (Section 1.6 of Chapter 1) may be analyzed in a similar manner.



**Table 5.4** Some major index headings in CAPLUS in the field of molecular biology

| Index heading | Valid periods | Number of records | Some alternative terms, notes |
|---|---|---|---|
| Chromosome | 1972– | >52 000 | Used for Genetic maps; more than 80 narrower and related terms |
| Deoxyribonucleic acids | 1972–1996 | >131 000 | Replaces Nucleic acids, deoxyribo-; new term: DNA (>40 000) |
| DNA sequences | 1997– | >40 000 | Old term: Deoxyribonucleic acid sequences (>90 000); related term: cDNA sequences (>34 000) |
| Gene | 1972–1996, 1999– | >240 000 | Used for general studies on genes |
| Gene and genetic element | 1982–1991 | >113 000 | New term: Gene (>240 000); new term: Genetic element (>48 000) |
| Gene, animal | 1992–1996, 1999– | >140 000 | Old term: Genes (animal) |
| Gene, plant | 1992–1996, 1999– | >16 000 | |
| Genes (microbial) | 1997–1999 | >21 000 | New term: Gene, microbial |
| Genetic element | 1992–1996, 1999– | >48 000 | Old term: Gene and genetic element |
| Genetic engineering | 1982– | >9000 | Narrower term: Molecular cloning (>57 000) |
| Genetic methods | 1992– | >8000 | Narrower terms include Genetic mapping (>30 000), Genetic vectors (>10 000), PCR (>16 000), Nucleic acid hybridization (>15 000) |
| Genetics | 1967– | >43 000 | |
| Molecular cloning | 1977– | >57 000 | Used for Cloning, gene fusion |
| Mutation | 1967– | >86 000 | |
| Nucleic acids | 1967– | >35 000 | The general heading has more than 150 narrower and related terms |
| Nucleosides | 1967– | >15 000 | >60 narrower and related terms |
| Nucleotides | 1967– | >36 000 | |
| Plasmids | 1967–1971, 1997– | >6000 | Term used 1972–1996: Plasmid and episome (>36 000) |

(*continued overleaf*)

**Table 5.4** *(continued)*

| Index heading | Valid periods | Number of records | Some alternative terms, notes |
|---|---|---|---|
| Promoter (genetic element) | 1997– | >21 000 | |
| Ribonucleic acids | 1972–1996 | >125 000 | New term: RNA (>8000) |
| Ribonucleic acids, messenger | 1972–1996 | >60 000 | New term: mRNA |
| Ribonucleic acids, ribosomal | 1972–1996 | >14 000 | New term: rRNA |
| Ribonucleic acids, transfer | 1972–1996 | >13 000 | New term: tRNA |
| Ribonucleic acids, viral | 1972–1996 | >11 000 | New term: viral RNA |
| Transcription factors | 1997– | >33 000 | >120 narrower and related terms |
| Transcription, genetic | 1992–1996, 1999– | >29 000 | Transcription |
| Transcriptional regulation | 1998– | >14 000 | |

number of records with the various headings, and the last column indicates *some* alternative terms and notes. Thus, the index heading 'Deoxyribonucleic acids' was used for general studies on this class of substances between 1972 and 1996, and just over 131 000 records have this heading. Since 1996, the heading 'DNA' has been used, and it is already appearing in more than 40 000 records. Similarly, the heading 'Ribonucleic acids' and its various subsets have been replaced by 'RNA', 'mRNA', and so forth.

Often, within these general headings are numerous narrower headings that may be even more commonly listed. For example, 'Genetic methods' appears in just over 8000 records, whereas the narrower headings 'Genetic mapping', 'Genetic vectors', 'PCR', and 'Nucleic acid hybridization' each appear in well over 10 000 records. The indexing principle is that precise headings are applied, except for very general studies that are indexed at the broader level.

It will be apparent from Table 5.4 that some of the changes in indexing will be of little consequence to the user of SciFinder Scholar. Thus, an **Explore by Research Topic** 'I am interested in ''microbial genes'' ' will retrieve 'gene, microbial' and 'genes (microbial)'. However, will an explore 'I am interested in ''DNA sequences'' ' include 'deoxyribonucleic acid sequences' or 'cDNA sequences' as search terms? And will an explore 'I am interested in ''mRNA'' ' include 'messenger ribonucleic acids' in the search? The answer is simple: 'Just try it!' (Exercise 5.8 explores this issue). As usual, it helps if the scientist applies 'scientific method', which in the first instance means that the searcher has a general understanding of the field (author- and indexer-related terminologies), next tries an initial experiment, then carefully observes the result, and finally modifies the search to best meet the requirements.



One point in favor of searching for nucleic acids and derivatives is that there are very few building blocks (e.g. there are only a few nucleic acid bases) and relatively few functions (e.g. most simply provide the code for nature to manufacture proteins). Further, molecular biologists have developed simple ways to describe 'substances', for example, the isoleucine valine genes and operons (and related proteins) have letters or numbers to denote specific types (*ilvA*, *ilvB*, *ilv*1, *ilv*3, *ilvGMEDA*, etc.). Generally, these terms are very specific, and searches through **Explore by Research Topic** and then choice of candidate 'as entered' or the 'concept' will quickly lead the scientist to precise answers.

The same cannot be said about proteins, and particularly about enzymes for which numerous different terms are used. Of the two different approaches mentioned in Section 5.2, the option starting with **Explore by Chemical Substance** will give precise substances, but many CAS Registry Numbers may need to be searched to cover the alternatives. For example, when 'tyrosine kinase' is searched through Scholar in REGISTRY, the enzyme (CAS Registry Number 80449-02-1) is retrieved, and CAPLUS contains more than 7400 references with this CAS Registry Number. However, there are more than 1000 records in REGISTRY that refer to 'tyrosine kinase' (different isoforms, genes, receptors, clones, different animal and plant sources of the enzyme, etc.).

Meanwhile, the option starting with **Explore by Research Topic** may be more comprehensive but may lack precision. For example, a search on 'tyrosine kinase' gives more than 48 000 records 'as entered' of which about half are in CAPLUS. Now, in addition to all the tyrosine kinase receptors, genes, sources, and so forth, answers also contain *inter alia* tyrosine kinase inhibitors. The solution is to start at more general levels, and then to use all the analyze and refine tools available in SciFinder Scholar to increase precision.[2]

## 5.5 Searching for Information on Polymers

Searching for information on synthetic polymers also presents its challenges! Although the basic registration starts from the monomer components (Appendix 5.4), various polymer forms (block polymers, graft polymers, etc.) and blends have slightly different registrations. An additional concept is that

[2] Additionally, SciFinder users have BLAST® search options (Section 4.7 of Chapter 4).

of structurally repeating units (SRUs), which are applied to those polymers of clearly defined composition. A discussion of some of the issues that apply to siloxane polymers is available at *http://www.cas.org/silox.html*, and this discussion may be used by advanced searchers as a basis for understanding the registrations of other polymers.

An excellent discussion of opportunities for searching for polymers through SciFinder Scholar is available through *http://www.cas.org/SCIFINDER/SCHOLAR/SOLUTIONS/polymer1.html*, and the entry screen is shown in Figure 5.5. Links from this page describe details of searching by formulas, structures, and keywords.

While any of these options may be used to obtain the required level of precision in answers, the ability to get related substances adds another dimension and greatly facilitates searching for information on polymers and

Address http://www.cas.org/SCIFINDER/SCHOLAR/SOLUTIONS/polymer1.html

**Preview** With SciFinder Scholar, you can locate polymer information matching

- A specific polymer name or CAS RN
- Structures of specific monomers
- Keywords describing general classes of polymers

**Decision table** SciFinder Scholar offers a number of pathways to locate polymer information. Choose a pathway that will yield results in line with your information needs.

| When you want | And you know | Use this SciFinder Scholar strategy |
|---|---|---|
| Precise information | • Polymer tradename, acronym, or company designator<br>• CAS RN | *Explore by Substance Identifier* |
| General information | The molecular formulas for all the monomeric units | Polymer Formula Searching |
| | The structure(s) for some or all of the monomeric units | Polymer Structure Searching |
| Very general information | Terms describing the general class of polymers | Polymer Class Searching |

**Figure 5.5** 'SciFinder Scholar Solutions' helps users with searching for polymers.



the identification of relevant CAS Registry Numbers. For example, there are many hundreds of 'nylons' (e.g. nylon 11, nylon 6/66, etc.) in REGISTRY and to find CAS Registry Numbers by name searches would be tedious. Further, if general structure units (e.g. $-N(CH_2)_5N-$) are of interest, then even after the use of 'additional options' in structure searching to restrict answers to polymers, the query may be too general to complete because of system limits.

However, the scientist now has the option of starting the search from **Explore by Research Topic**, then working through to substances through **Get Related: Substances**, and the only restriction is that bibliographic answer sets need to contain fewer than 1000 references. One of the many ways around this is to search initially for the property of interest, so, for example, the search may start with **Explore by Research Topic** 'I am interested in ''hydrogen bonding or H bonding in nylons'' '. This gives more than 450 references in which the concepts are closely associated and **Get Related: Substances** gives more than 1000 substances, which may be refined by the structure fragment $-N(CH_2)_5N-$. A substructure search now gives around 20 substances including nylon 55, nylon 56, nylon 57, and so on, and includes polymers that are registered as SRUs and are monomer-based.

## 5.6 Exercises

5.1 Using the two different approaches in Section 5.2, compare answers that discuss Lorenzo's oil. Locate a publication by Lorenzo's parents (Last name: Odone) and use **Get Related**: **eScience** to find information about the film made about the case.

5.2 Using the two different approaches in Chapter 5.2, find reports that discuss urocanic acid in relation to skin cancer.

5.3 Find studies on gene sequences and HIV or AIDS. How many records are in CAPLUS and how many in MEDLINE? How many records are patents? How many of these have the index heading 'molecular cloning' and how many have English as the original language of publication?

5.4 A medicinal scientist interested in photodynamic therapy for psoriasis ran an **Explore by Research Topic** 'I am interested in ''psoriasis with photodynamic or PDT'' '. However, the answer set in which the concepts were 'anywhere in the reference' showed 150 records in CAPLUS and only 32 in MEDLINE. There clearly is something wrong

with the low numbers of hits in MEDLINE. How would you approach the problem?

5.5 Find the substance: sumatriptan. From the substance screen get references:

5.5.1 all references;

5.5.2 on the adverse effects;

5.5.3 on Analytical Study;

5.5.4 on Spectral Properties; and

5.5.5 on Uses.

In each case, compare the numbers of records in CAPLUS and in MEDLINE, and comment on the reasons (i.e. type of search terms used) for the retrieval of the various answer sets.

5.6 Comment on the indexing of

5.6.1 nylon 610 (CAS Registry Number 9008-66-6) and nylon 610, monomer-based (CAS Registry Number 9011-52-3) and

5.6.2 Ube nylon 3024B (CAS Registry Number 25038-74-8) and nylon 12 (CAS Registry Number 24937-16-4).

5.7 In seeking polysiloxanes that are heat resistant, a scientist first undertakes **Explore by Research Topic** 'I am interested in ''heat resistance of siloxanes or polysiloxanes'' '. The answers 'closely associated' are chosen and following **Analyze References** by **Index Term**, the answers are narrowed to those indexed with heading 'Heat resistant materials'. From this point, how may the actual polysiloxanes be retrieved? (Your answer should include around 250 substances!).

5.8 Determine whether the following are searched within the same 'concept' in CAPLUS:

| | | |
|---|---|---|
| 5.8.1 messenger ribonucleic acid | mRNA; | |
| 5.8.2 deoxyribonucleic acid sequences | DNA sequences | cDNA sequences; |
| 5.8.3 protein sequences | amino acid sequences | peptide sequences. |

5.9 Locate index terms for sequence data in MEDLINE.



5.10 Find information on studies on DNA sequences in the tsetse fly.

5.11 Predict the 'concepts' searched in an **Explore by Research Topic** 'I am interested in ''hepatitis C with immunoglobulin A'' '. How may records only to immunoglobulin A be obtained? Identify the index headings used in CAPLUS and in MEDLINE for immunoglobulin A, and find records that report immunoglobulins A and hepatitis C.

Chapter 6

# Searching for Chemical Reactions

## 6.1 Introduction

There are numerous ways in which reactions are described and numerous reasons for which they are performed. Reactions may be described by their type (oxidation, reduction, addition, elimination, substitution, polymerization, cyclization, metathesis, etc.) or by their intent (reaction, preparation, mechanism, etc.). Chemists also describe reactions by their names (Wittig reaction, Dess–Martin oxidation, hydroboration, Grob fragmentation, etc.) or by acronyms (IMDA – intramolecular Diels–Alder). Substances involved may be described as reactants, reagents, starting materials, products, or catalysts. Other factors relating to reactions are conditions like solvents, temperature, photochemical reactions, and yields.

The various terms that describe reactions are used extensively by *authors* in titles and abstracts, and a *wealth of chemical reaction information appears in these fields in CAPLUS.* Chemical reaction information is also entered by *indexers* in CAPLUS, primarily as index headings and through the Chemical Abstracts Service (CAS) Roles (Section 1.5.4 of Chapter 1 and Appendix 2) associated with CAS Registry Numbers or substance class headings. Additionally, CAS has created a separate chemical reaction database (CASREACT®, Section 1.10 of Chapter 1) in which reactions are fully indexed with atom-by-atom mapping, bonds formed or broken in the reaction, and a number of other identifying factors including, *inter alia,* type of reaction and yield. *There are a very large number of ways to describe, and hence also to find information on, reactions.*

Although authors may enter reaction information in titles and abstracts, the reaction information may often appear only in the text of the full article.

Consider, for example, the bibliography and abstract in the article by Zhang and Breslow (Appendix 7) and the series of reactions in Schemes 1 and 2. These reactions are not mentioned in the title or the abstract, but Schemes 1 and 2 are indexed both in CAPLUS and CASREACT. For example, Figure 6.1 shows the sequence in Scheme 1 (Appendix 7) together with the CAS Registry Numbers of the substances involved and the roles associated with these Registry Numbers in the bibliographic database. In the record, 42 CAS Registry Numbers are indexed with the role PREP and 47 CAS Registry Numbers with the role RCT. For details of the indexing, the record (Accession Number 126:183046) in SciFinder Scholar may be viewed. Scheme 2 is also fully indexed in a similar way in this database.

On the other hand, the record for the original article in the *reaction database* is quite different. Of particular note is that this database contains only information on Scheme 2 (there are a number of indexing policies applied, and in particular, only key reactions are entered in the reaction database), and the six synthetic steps from 2-bromopyridine (**11**) to the substituted cyclodextrin

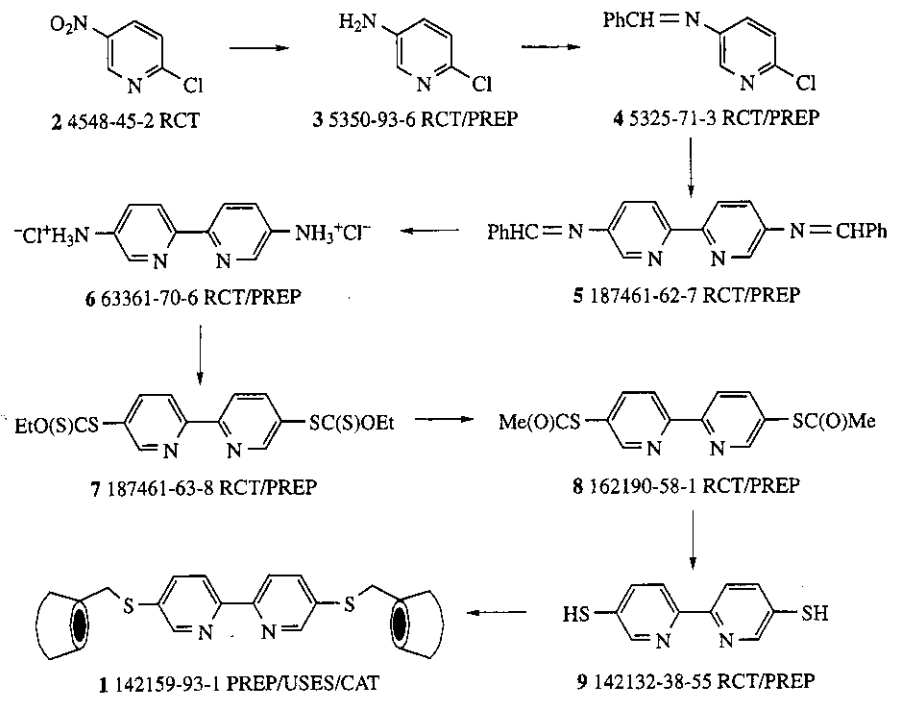



**Figure 6.1** Indexing of reaction sequence in Scheme 1 (Appendix 7). (Substance numbers correlate with those in the original publication.)



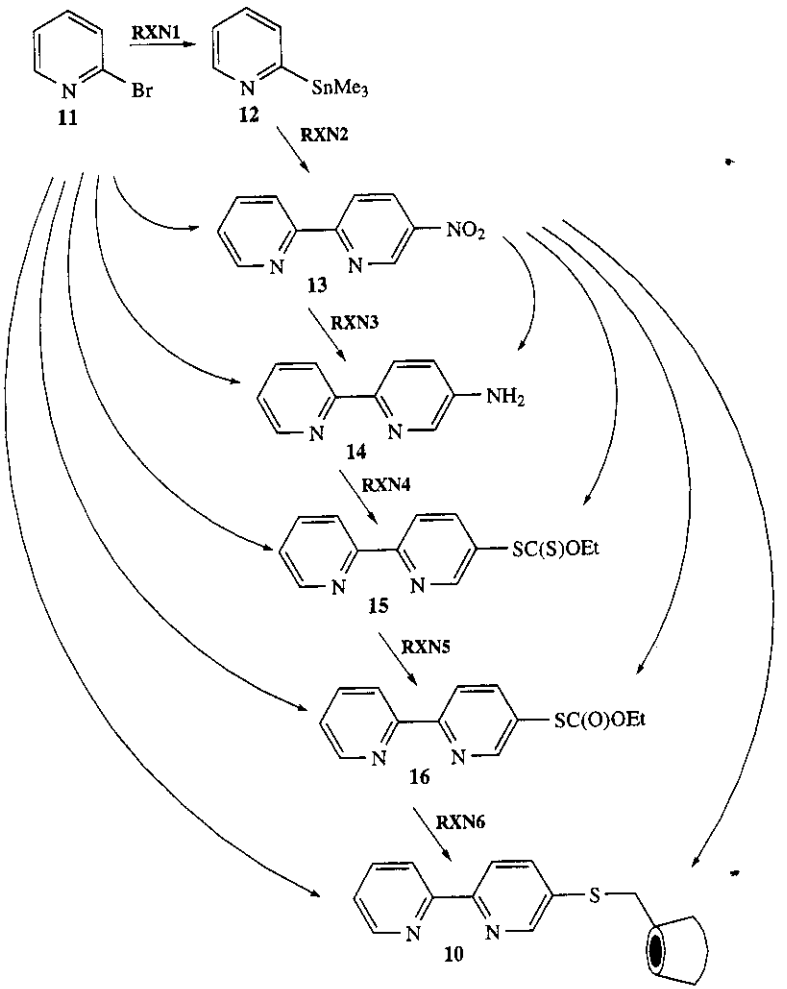



**Figure 6.2** The six reactions in Scheme 2 (Appendix 7) are indexed as 21 'reactions' in CASREACT in order to allow for searches on multistep reactions.

(**10**) are indexed as six different reactions (Figure 6.2). However, in order to allow for questions that involve multistep reactions, 21 'reactions' appear in the record in the database for this sequence. That is, the two-step sequence from (**11**) to (**13**) and each of the other multistep reactions from (**11**) up to the six-step sequence (**11**) to (**10**) are indexed as 'reactions', together with the various multistep processes from synthetic intermediates (**12**) through (**16**).

**Table 6.1** CAS databases searchable for reaction information in SciFinder Scholar

| Search option through SciFinder Scholar | Database searched | Coverage |
|---|---|---|
| **Explore by Research Topic or Refine by Research Topic** | Words in title, abstract, and indexing in CAPLUS | More than 8000 serials and patent documents from more than 30 countries. Coverage from 1907– |
| **Explore by Chemical Substance** (Chemical Structure), then **Get Substances,** followed by **Preparation** or **Reactant** | Substances are found in REGISTRY, then answers in CAPLUS have CAS Registry Numbers closely associated with the PREP or RCT roles | REGISTRY commenced in 1965, but some substances reported earlier are included |
| **Explore by Chemical Substance** (Chemical Structure) then **Get Reactions** | CASREACT | Selected journals in synthetic organic chemistry 1907–; some patents 1991– |

It is apparent from the preceding example that there are many fundamental issues to address when setting up the reaction query. First, the CAPLUS and CASREACT databases are quite different, particularly with respect to the number of journals covered and time. Table 6.1 summarizes the options together with the ways in which the databases may be searched through SciFinder Scholar.

Second, the bibliographic database is by far the more comprehensive since it covers more than 22 million records from 1907. Searches in the title and abstract fields may lack precision, but more precise information on reactions and preparations may be retrieved by searching CAS Registry Numbers and linking them with CAS Roles. Currently, the PREP and RCT roles occur in more than 2.9 and 2.4 million of the records, respectively, and since these may appear many times in a single record, the opportunities to find preparation and reaction information are very extensive.

Finally, searches in the reaction database may be refined with great precision, but they may lack recall because fewer original articles have records in the reaction database and because of the indexing issues explained earlier.

## 6.2 Specific Search Options in CASREACT

The structure drawing screen (Figure 3.2) has five icons in the vertical palette that are used specifically for searching for reactions in CASREACT. Two



of them (**Get Reactions: Bonds to be formed or broken in a reaction** and **Map atoms in a reaction**) relate to special concepts associated with reactions. For example, if the user requires references to reactions of the type shown in Scheme 1 (Figure 6.3, i.e. perhaps Wittig type reactions involving $Ph_3P = CHOMe$), one option would be to draw the scheme exactly as shown and then to choose **Get Reactions**. SciFinder Scholar looks for all reactions in CASREACT that have the cyclohexanone substructure as a reactant and the product as a part structure. Many answers will be of the required type (e.g. Reaction A), but Reactions B, C, and D will also be retrieved since they meet the search requirements. However, these last three reactions are of quite a different type to the intended Wittig reaction.

There are many ways to proceed. One would be to request that the double bond in the starting material be 'formed or broken', when reactions of types A and B only would be retrieved. Another would be to request that the double bond in the product be 'formed or broken' (reactions of types A and C only would be retrieved), whereas if the double bonds in both reactant and product were tagged, reactions of type A only would be retrieved. Alternatively, **Get Reactions: Map atoms in a reaction** may be used to ensure greater precision. The search process in SciFinder Scholar is quite precise and works very differently from other commonly used reaction databases, which often require

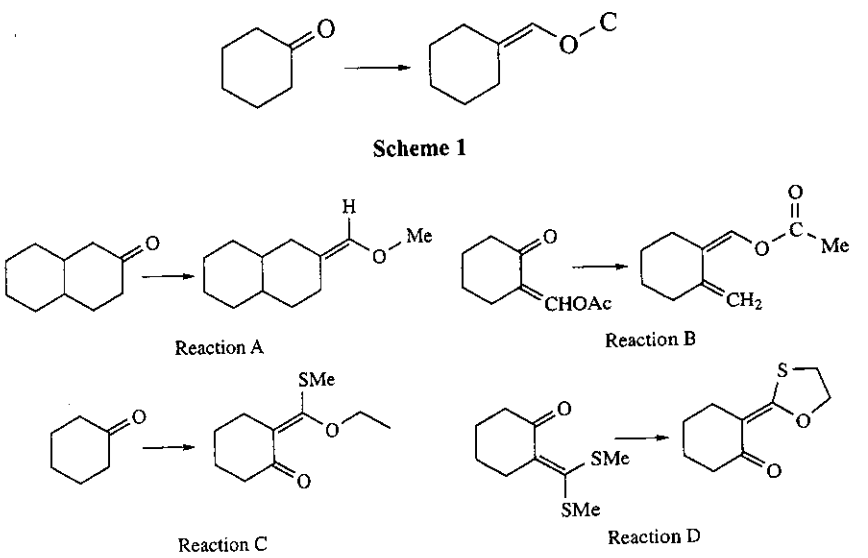



**Figure 6.3** Examples of actual answers found in the search on the reaction shown in Scheme 1.

extensive mapping of atoms/bonds to obtain reasonably precise answers. In SciFinder Scholar, it generally is sufficient to map only one bond to be formed or broken, and atom mapping, if used, should be kept to a minimum.

Another icon **Get Reactions, A → B (add a reaction role)** is used to specify whether the structure drawn should be in the starting material, in the product, or anywhere in the reaction. However, note that when the icon **Get Reactions, → (draw a reaction arrow)** is used, the reaction roles are applied automatically. The final icon, **Functional Groups**, is used to specify functional groups. At this stage, it is sufficient to know that these five options exist. Their applications are discussed in sections that follow in this chapter.

## 6.3 Reaction Search Strategies

As usual, there are a number of questions to consider before the search is conducted. For example, one of the key reactions in the article by Zhang and Breslow is the conversion in Figure 6.4, and the issue is how to proceed to find such a process. The first point to note is that there is no mention of anything relating to this reaction in the title or in the abstract, so on this occasion, searches on words in these fields will not retrieve this record. The searcher then must rely on terms entered by the indexer. Since information on substances is entered through their CAS Registry Numbers, the challenge is to find the appropriate Registry Numbers, and it helps if the user first considers a number of questions including:

(1) Is only this exact reaction required?

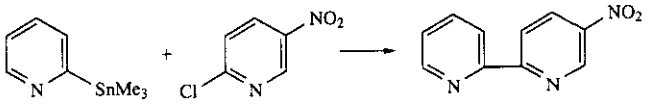
(2) Would some variations be acceptable, for example, would a preparation from the pyridine 2-tri*butyl*stannane and 2-*bromo*-5-nitropyridine be acceptable?

(3) If the exact reaction, or any variations like those in (2) above, are not known, then what type of information may be acceptable? Would, for example, information on the preparation of a possible precursor (e.g. the aminobipyridine rather than the nitrobipyridine) be acceptable, since a subsequent conversion may be possible to give the new substance?

(4) Does a crucial aspect of the question really relate to finding reactions in which a bond is formed between two pyridine rings in the $\alpha$-position?

(5) If attempts to answer the above questions in the reaction database do not produce the type of information required, then how may the problem be solved in the bibliographic database?






13737-05-8 4548-45-2 14805-00-6

**Figure 6.4** Stille reaction from the Zhang and Breslow publication (Appendix 7).

It really does not matter whether the searcher *first* approaches the problem from the most specific option (searching the specific reaction in CASREACT) or from the most general option (searching CAPLUS for preparations of products or reactions involving starting materials). What is more important is that the searcher realizes that there are a number of different approaches and that several of them are attempted. More general searches are not only more likely to retrieve specific processes but also are likely to retrieve alternatives that the searcher may not have considered initially.

### *6.3.1 Get Substances and Get Reactions*

**Get Substances**

At a reasonably general level, the structure of the nitrobipyridine (14805-00-6) may be built, and a full substructure search gives all the substances with the nitrobipyridine part structure. When **Get References** is chosen (followed by **Only of the following type: Prep**), records that have any of the CAS Registry Numbers from the substructure search closely associated with the CAS Role PREP will be retrieved in CAPLUS. Such a search currently gives about 45 substances, which lead to about 13 references that describe preparations.

If the user is interested only in such nitrobipyridines, then this answer set may meet the requirements. Certainly, because there are very few answers in this broader database, the user probably would not need to proceed to CASREACT, which would provide fewer answers. However, the chances are that the user may also be interested in related derivatives, in which case, an even more general search may need to be undertaken.

There are many ways to proceed with either more general or more specific questions, and a summary of some of the options is given in Table 6.2. The intention of Entry 1 (Table 6.2) is to find all the bipyridines and then limit references to those that describe their preparations, but as such, a general query does not complete, and the 'autofix' option is required. While this excludes all fused rings, still more than 7800 substances are retrieved. The search now probably is too general.

**Table 6.2** Search options in the substance and reaction databases in SciFinder Scholar

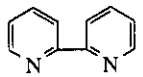
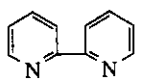
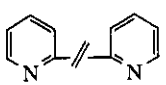
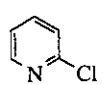
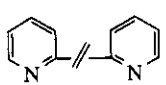
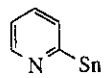
| Entry | Search query | Search request | What SciFinder Scholar does | Number of answers |
|---|---|---|---|---|
| 1 |  | 1. **Get Substances**<br>2. Substructure<br>3. **Get** (all) **References** or<br>4. **Only of the following type: Preparation** (or **Reactant**) | *Steps 1&2:* a substructure search of the query in REGISTRY[a]<br>*Step 3* searches CAS Registry Numbers from the substructure search in CAPLUS<br>*Step 4* searches all records in CAPLUS with CAS Registry Numbers closely associated with CAS Role PREP (or RCT) | More than 7800 substances with more than 2800 references to their preparation[a] |
| 2 |  <br>Product | 1. **Get Reactions** | Searches all records in CASREACT with CAS Registry Numbers (mapped with role PREP) from a substructure search | More than 1100 records. When **Refine Reactions** is chosen and the structure Sn[b] drawn as a reactant, there are more than 70 records |
| 3 |  <br>Product | 1. **Get Reactions** | Similar search to Entry 2, except that only answers with the tagged bond formed in the preparation are retrieved | More than 190 records. When **Refine Reactions** is chosen and structure Sn[b] drawn as a reactant, there are more than 20 records |
| 4 |  <br>Reactant<br>  <br>Product | 1. **Get Reactions** | Similar search to Entry 3, except that only answers with the 2-chloropyridine substructure as reactant are retrieved | More than 40 records. When X (any halogen) is substituted for Cl, there are more than 100 records |



**Table 6.2** (*continued*)

| Entry | Search query | Search request | What SciFinder Scholar does | Number of answers |
|---|---|---|---|---|
| 5 |  <br>Reactant<br>+<br>  <br>Reactant | 1. **Get Reactions** | Searches all records in CASREACT with CAS Registry Numbers from substructure searches mapped with role RCT | More than 20 records |

Note: Number of records in CASREACT correspond with the number of references, and a single reference may contain a number of examples of the reaction.
[a] The search query was too general and 'autofix' was required for the substructure search to run to completion.
[b] **Refine Reactions** returns the user to the structure drawing screen. Drawing a node Sn and specifying the node as reactant or reagent was chosen in this example in an attempt to find information on Stille-type couplings.

### Get Reactions

A more specific search is shown in Entry 2 (Table 6.2). Since any search in CASREACT automatically applies a screen to limit answers to those with hits in this database, it is now found that it is not necessary to apply the ring locking tool. Around 1100 references are retrieved and one is shown in Figure 6.5. The number of answers actually is the number of *references* in CAPLUS, and a single reference may have a number of different hit reactions.[1] For example, the reference *Tetrahedron*, 2000, **56**, 3575 in Figure 6.5 has 15 hit reactions, although of course some of these may be 'hits' from multistep reactions. Depending on the real search intentions, this may be a useful answer set, although probably the scientist still would not be interested in so many answers.

Even more precise answers may be obtained by first tagging the bonds to be formed in the preparation of the bipyridine or by specifying the starting materials. Entries 3 and 4 in Table 6.2 show examples and give the number of answers received in each case. While just over 40 references are retrieved in the reaction search under Entry 4, many more answers are retrieved when the generic group for halogens (the symbol, X) is used in the query. Finally, an even more specific query is represented in Entry 5 and it gives more than 20 records.

Accordingly, the query in Figure 6.4 may be approached in many ways! The issue is for the scientist to be aware of all the opportunities and to use one or more strategies as appropriate.

[1] SciFinder Scholar now lists all hit reactions.

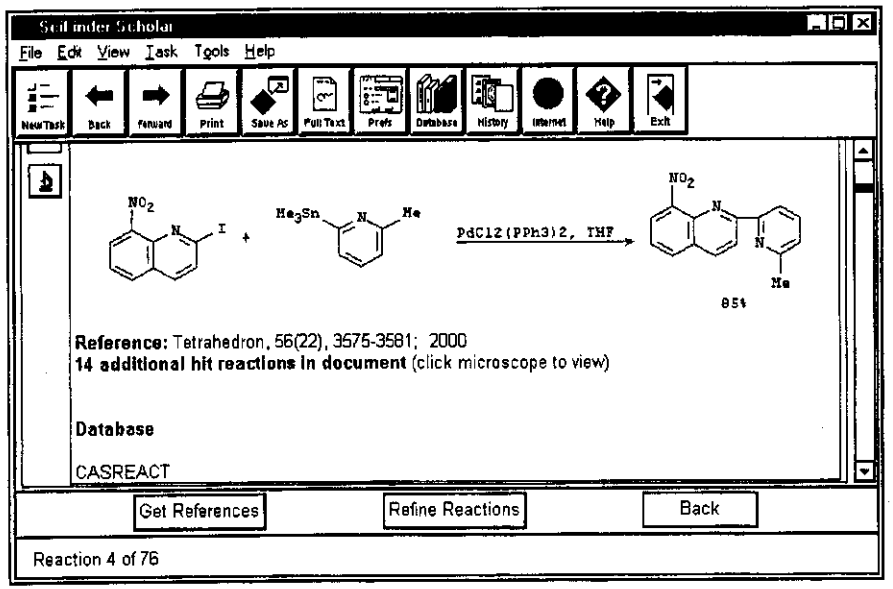



**Figure 6.5** Example of answer from Entry 2 (Table 6.2). Copyright the American Chemical Society and reproduced with permission.

### Refine Reactions

Once an answer set has been obtained, the usual SciFinder Scholar options of print/save, the microscope, and **Get References** may be followed. However, as seen in Figure 6.5, there is the additional option to **Refine Reactions**, which gives four different options (Figure 6.6), any of which may be followed. Under the different options are submenus that guide the searcher. Care needs to be exercised in their use (e.g. *inter alia*, yields of products may not have been specified in all the original articles). When it is critical to use these options, and if the searcher is uncertain about how precise or comprehensive searches are, then the searcher should check with the account key coordinator.

When **Chemical Structure** is chosen, the user is returned to the structure drawing screen and any structure with any specified role or mapping option or any functional group may be drawn. The new search then operates on the answer set already obtained. For example, when the initial answer set in Entry 3 (Table 6.2) is refined with the 'structure' of a Sn atom as a reactant, 22 references are obtained and all involve Stille couplings of the type in Figure 6.4.



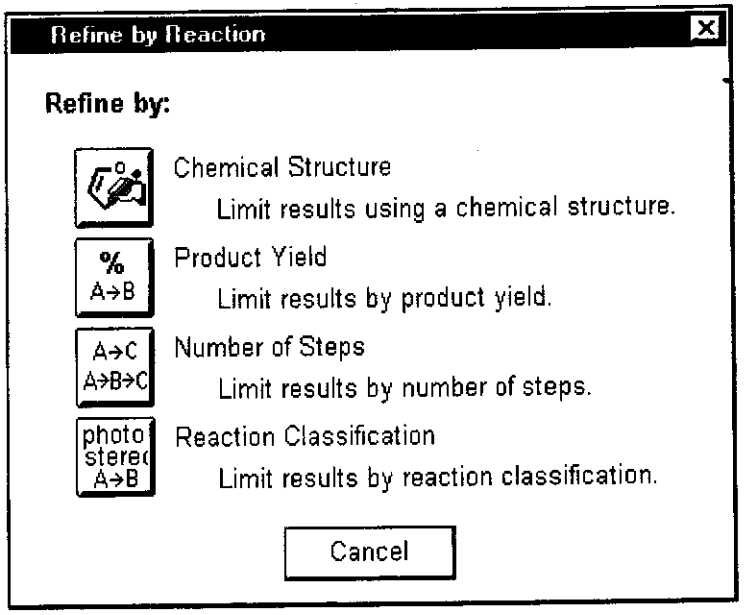



**Figure 6.6** The four choices when **Refine Reactions** is chosen. SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

## 6.3.2 *Using Functional Groups*

Although chemists often are interested in the reactions of substances of a certain type, sometimes the key requirement is simply the chemistry of a particular functional group. Of course, functional groups may always be represented by part structures, and one approach to find the methods to convert 1,2-diols (glycols) to aldehydes may be to search for the reaction in Figure 6.7.

To do this, SciFinder Scholar would have to find all reactions that have the glycol substructure in the starting material and the aldehyde substructure in the product, and such a general search would easily exceed the system search limits even in CASREACT. The solution is to search by functional groups.

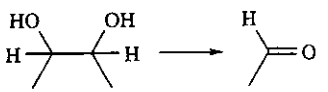



**Figure 6.7** Structure query for search for methods to convert glycols to aldehydes.

A large number of functional groups are built into SciFinder Scholar and these are displayed by clicking on the functional group icon in the structure drawing screen (Figure 6.8). Clicking on the functional group (left column) displays more specific functional groups in the class (right column). Searches may be performed at the specific or class level, and further restrictions may be made to require the functional groups to be in rings or chains. Once the required group has been chosen, clicking on the structure drawing screen enters the functional group term.

In this way, the query in Figure 6.9 may be set up adding the reaction arrow (→) or the reaction role icon (**A→B**). Actually, when the latter icon is clicked on a functional group in the structure drawing screen, there are four choices (reactant, product, any role, and non-reacting) that may be used as required.

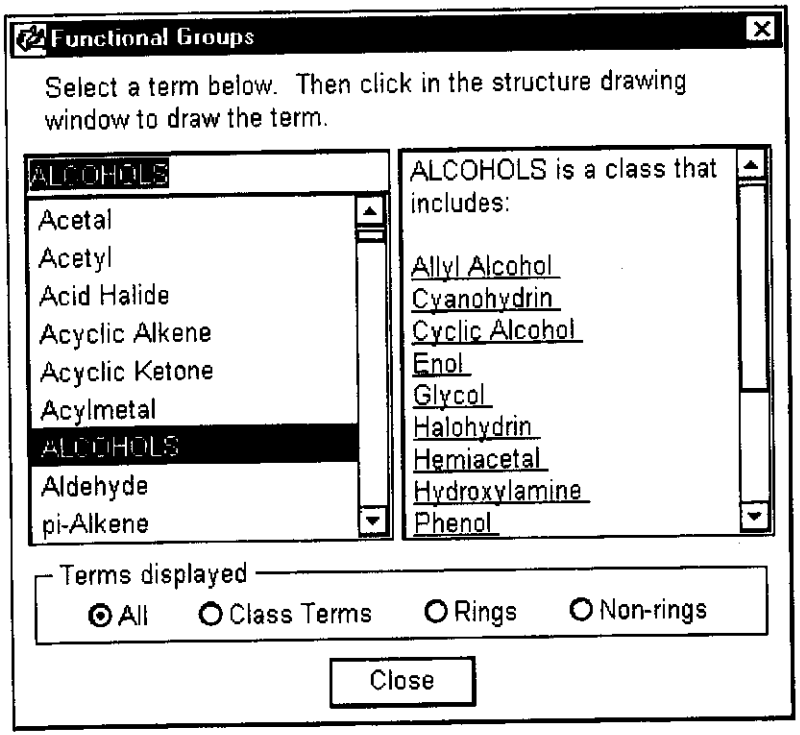



**Figure 6.8** More than 100 functional groups are built into SciFinder Scholar and may be used in the initial reaction query or within **Refine Reactions**. SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.



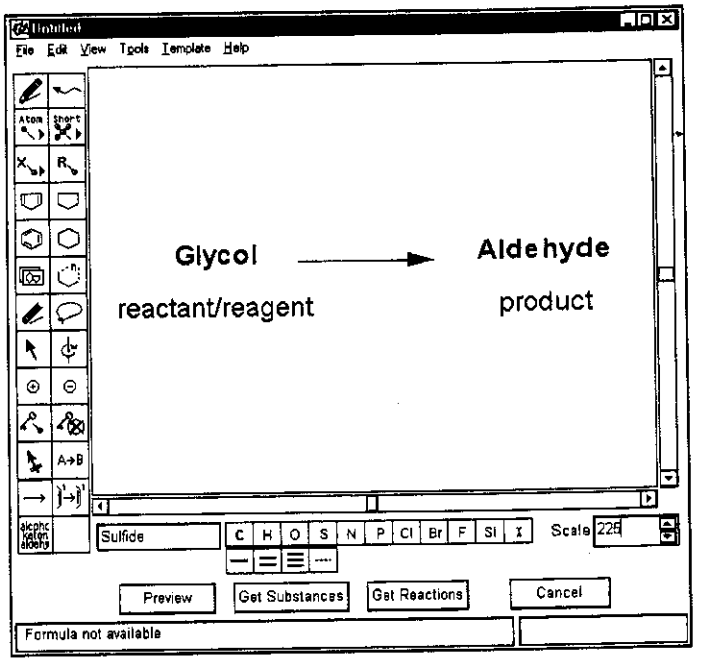



**Figure 6.9** Functional groups are entered in the structure drawing screen, and reaction roles may be assigned (through icon **A→B**). SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

When **Get Reactions** is selected, the result is an answer set of around 1000 reactions, which are shown in the format displayed in Figure 6.5.

A second important application of searches by functional group is the use of the 'non-reacting' option, which is particularly important when information about selective reactions of functional groups is required. For example, to search for reagents that will selectively oxidize sulfides to sulfones in the presence of other oxidizable groups (e.g. alcohols), it is easy to set up the functional group query: sulfides (reactant/reagent), sulfones (product), alcohols (non-reacting).

In summary, searching by functional groups helps overcome system limits for general searches and also allows for the very useful option of finding reactions that are selective for one type of functional group over another. Such searches may be conducted at any stage, that is, either as an initial search or under **Refine Reactions**.

### 6.3.3 Combining Functional Groups with Structures

The functional group and structure search options discussed in the preceding sections may be combined in any way to achieve different types of answers for different problems, and Figure 6.10 gives some options.

An initial search on the functional group query Figure 6.9 affords more than 900 answers, each of which may contain more than one example of the conversion. After **Refine Reactions** followed by **Refine by Chemical Structure** is chosen, a new query may be drawn, and Paths A, B, and C give typical examples of refinements that a chemist may require. Each of the answers in Figure 6.10 is from the reaction database, and from any display,

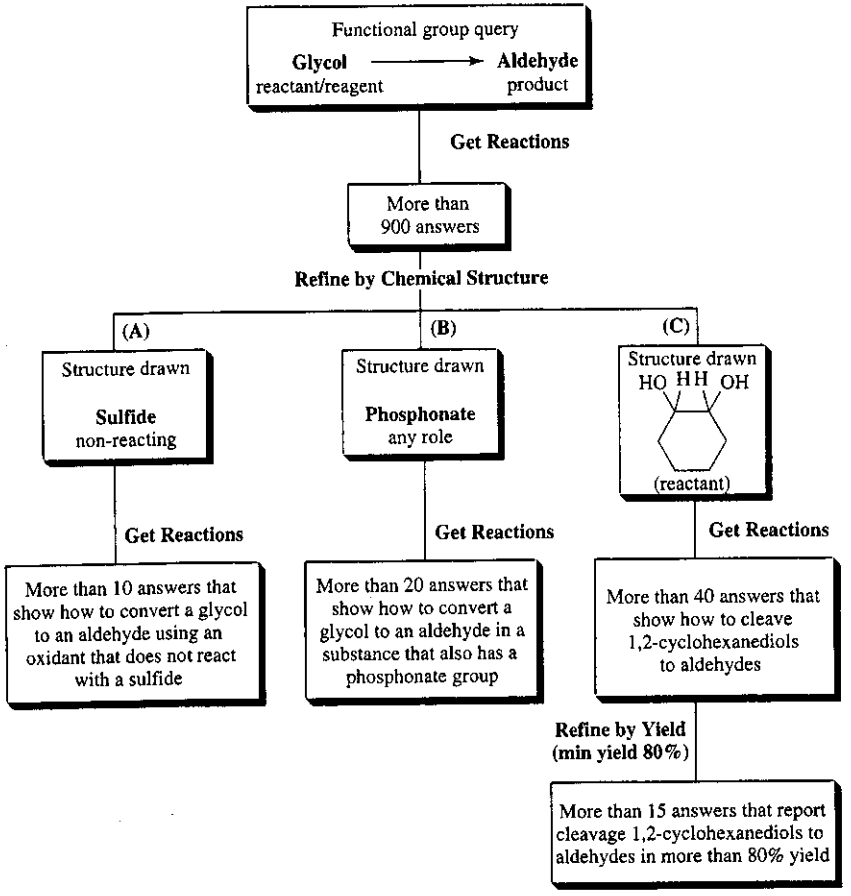



**Figure 6.10** Some of the opportunities to explore more precise reactions.

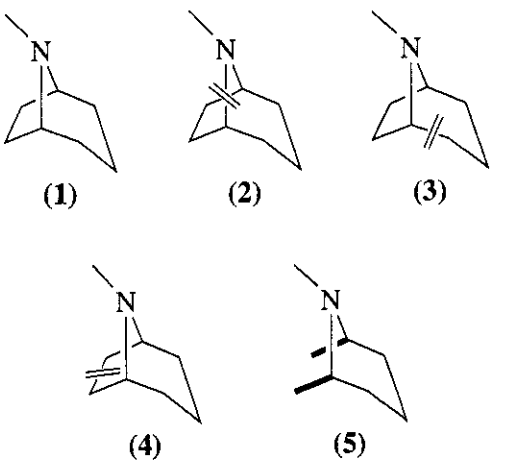


**Get References** may be chosen. Now the answers are from CAPLUS, and any of the options discussed in Chapter 2 may be used.

### 6.3.4 Retrosynthetic Analyses

To set up retrosynthetic analyses in CASREACT, it is simply necessary to draw the structure of interest, mark the bond that needs to be formed in the reaction, and specify the role 'product' for the structure. Different bonds may be marked in turn and the different synthetic routes may thus be evaluated.

For example, to perform retrosynthetic analyses on the tropane skeleton, the structure (**1**) is drawn, specified as a product, and the bonds are marked. In this particular case, when **Get Reactions** is chosen for structures (**2**) and (**3**), a number of the answers are the same because the well-known Mannich reactions involve formation of both types of bonds. When **Get Reactions** is chosen for structure (**4**), a number of the answers involve reduction of the corresponding alkene since currently SciFinder Scholar does not differentiate between bonds that are completely formed or broken and those that are partly formed or broken. To obtain answers of the former type, it is thus necessary to refine the initial answer set with the structure (**5**) and of course to ensure that the appropriate bonds are locked to chain values only!




The advantage of the analysis of this type is that the user is not making any commitment as to the nature of the starting material, and very often, new options become apparent. However, the user may subsequently refine answers with specific starting materials or through any of the other options shown in Figure 6.6.

**Table 6.3** Search options for reactions in CAPLUS

| Section of record | Origin of entry | How searched in SciFinder |
|---|---|---|
| Words in titles and abstracts; text-modifying phrases | Author[a] | Through **Explore by Research Topic** or **Refine by Research Topic** |
| CAS Chemical Reaction Index Headings | Indexer | Through **Analyze References** then **Index Term** |
| Supplementary Terms | Indexer | Through **Analyze References** then **Supplementary Term** |
| CAS Registry Numbers | Indexer | Preferably through **Explore by Chemical Substance** |
| CAS Roles | Indexer | Through **Explore by Chemical Substance** and then the list of roles that appears at **Get References** (Figure 3.21) |

[a]Titles and abstracts may be modified by the indexer for patents, and text-modifying phrases based on author terminology are entered by the indexer.

## 6.4 Searching for Reactions Through Explore by Research Topic

A record in CAPLUS has many points of entry into information on chemical reactions. These are summarized in Table 6.3.

**Explore by Research Topic** searches for words in titles, abstracts, and indexes, and there will be several occasions in which this option is preferred. For example, consider the options available if information on the reaction of ethyl propiolate with triethylamine is required (Figure 6.11). The most precise option is to draw the starting materials in the structure drawing screen, specify the structures as reactant or reagent, and then **Get Reactions**. If shortcut symbols or hydrogens on the carbons are specified, then any record retrieved in the reaction database will involve the precise reaction. However, in view of the time period searched and the fact that only representative reactions may be in CASREACT, the specific reaction may not be retrieved.

If hydrogens were not specified in the structure query, then substructure searches may be performed, but now the query contains very small structure fragments and the search may exceed system limits. Another approach may be to search functional groups, and a search for ALKYNE as a reactant or

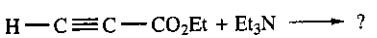


**Figure 6.11** What is the product formed immediately after triethylamine is mixed with ethyl propiolate?



reagent and TERTIARY AMINE as a reactant or reagent may be attempted. The searcher would need to consider carefully whether to specify the ester group since on chemical grounds it may or may not be involved in the reaction. Exercise 6.3 works through this approach.

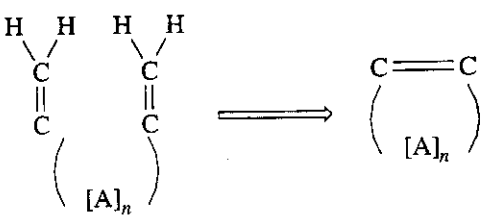
In the event that none of these options gives answers that really are satisfactory, the next option may be to consider structure searches in the substance database. Ideally, it would be necessary to find references that specified propiolic acid esters and tertiary amines as reactants, but currently it is not possible to combine answer sets like this in SciFinder Scholar. Further, system limits may arise for very general structure searches of this type.

Accordingly, probably the best initial option would be to enter relevant terms under **Explore by Research Topic**, and one option may be: 'I am interested in ''reaction of propiolate (propiolic) with triethylamine (trimethylamine, tertiary amine)'' '. When this is performed, around 20 references are obtained, and one of them actually is titled 'Reaction of propiolic acid esters with tertiary amines. Formation of betaines'!

## 6.5 Combining Structure, Reaction, Functional Group, and Keyword Terms

The integration of the world's largest chemical substance and bibliographic databases, with one of the world's largest chemical reaction databases, provides opportunities within SciFinder Scholar that are unique. The user merely has to be aware of the alternatives available and to work through the databases in a creative way.

To illustrate the opportunities, consider the ring closure metathesis reaction pioneered by Professor Robert Grubbs (*http://chemistry.caltech.edu/faculty/grubbs*) and illustrated in Scheme 1. Can SciFinder Scholar provide some insight into this reaction, perhaps particularly relating to the synthesis of medium size rings?



**Scheme 1** General scheme for ring closure metathesis reactions

If any linker between the reacting alkenes (e.g. chains with any atoms and with different lengths and types of bonds) is acceptable, the query may be as shown in Figure 6.12, but such a query will vastly exceed system limits. Accordingly, the functional group query (Figure 6.13) is searched and gives more than 14 000 answers that now may be refined with the query in Figure 6.12. This produces more than 1800 answers.

A quick glance at these answers indicates that the alkenes specified as reactants may be present as reactants in different steps of multistep processes, so the reactions are refined to single-step reactions (more than 1700 answers). Now, it is apparent that many of the answers involve Diels–Alder reactions of 1,3-butadiene, but to have apparently completely different reactions as matches for specific queries is not that uncommon!

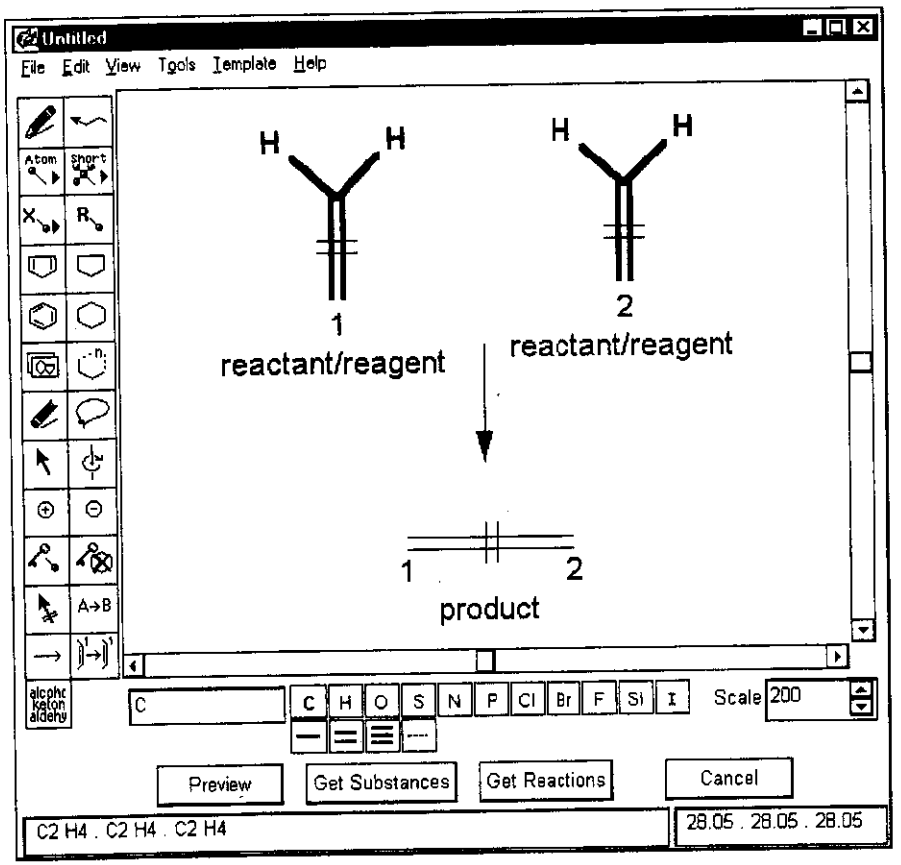



**Figure 6.12** Ring closure metathesis query. SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.



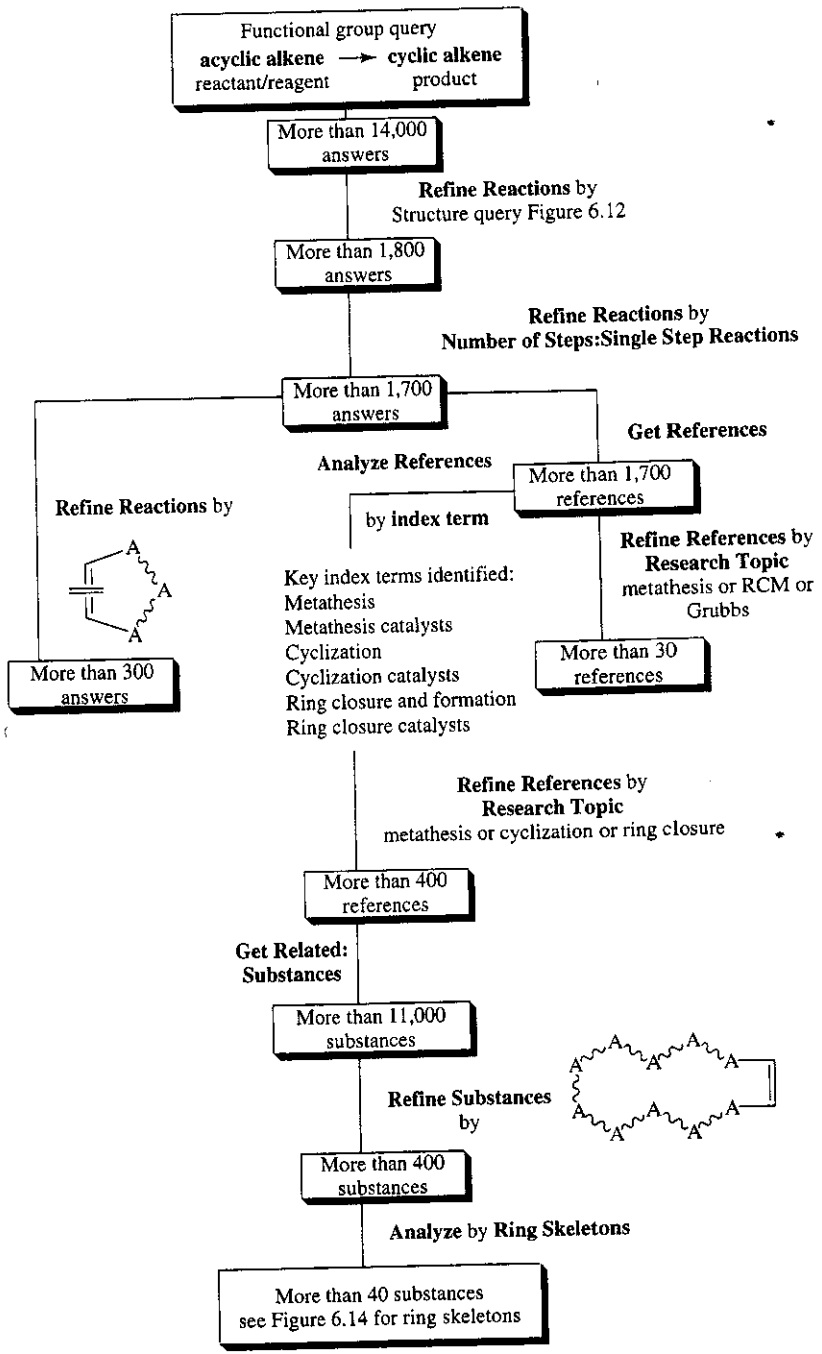



**Figure 6.13** Summary of searches for information on ring closure metatheses.

It is not that obvious how such answers may be removed by structure search queries. For example, one option, to refine reactions to those containing metals (since the required alkene metathesis require metal catalysts), is not appropriate since catalysts are not part of the reaction structure query. Another option, to go through all the reactions, is tedious, although in the long run it may prove the most effective. The usual approach is to use **Reaction Classification: Catalyzed**.

Another approach may be to **Refine Reactions** for the single-step reactions identified, and as indicated in Figure 6.13, a generic search on five-membered rings affords more than 300 answers. Many of these are relevant and show how five-membered ring compounds may be made from 1,6-dienes (including substances with heteroatoms) through ring closure metathesis reactions.

Similar generic searches on seven-membered rings also give quite precise answers, but the problem with Diels–Alder reactions occurs again when a generic six-membered ring query is tried. Further, if answers are refined with eight or larger ring systems, then only some answers are relevant since currently SciFinder Scholar includes two fused five-membered rings in answers for queries on eight-membered rings (unless the rings are isolated, and this would not be a good option in this case).

Next, approaches based on keywords may be considered, and after **Get References** is selected, the bibliographic records now obtained may be analyzed or refined in any of the ways outlined in Section 2.5.4 of Chapter 2. In particular, **Analyze References** by **Index Term** gives valuable information on terminology in the area, and a number of key terms become apparent (Figure 6.13). When these are chosen, an answer set of more than 400 references is obtained, and now only a few of these relate to Diels–Alder reactions (which have quite different index terms). As mentioned in Section 4.5.2 of Chapter 4, SciFinder Scholar and SciFinder subscription users now have the option to take all the indexed substances in CAPLUS records and to convert them to REGISTRY records (i.e. to a substance answer set of the type shown in Figure 3.8). All the *substance* analyze and refine options are now available, and either from the initial answer set of substances or from subsets refined by structures, the various ring skeletons may be identified.

For example, after the over 400 records mentioned above are obtained, **Get Related** then **Substances** is chosen and the REGISTRY answer set contains more than 11 000 substances. When these are refined with the 12-membered ring query (Figure 6.13), more than 400 substances are retrieved but some of these have 6 : 6 : 4 fused ring systems (since these have 12 atoms in the outer ring). However, 12-membered rings may be found through **Analyze** by **Ring skeletons**, and some of the ring skeletons retrieved are shown in Figure 6.14.



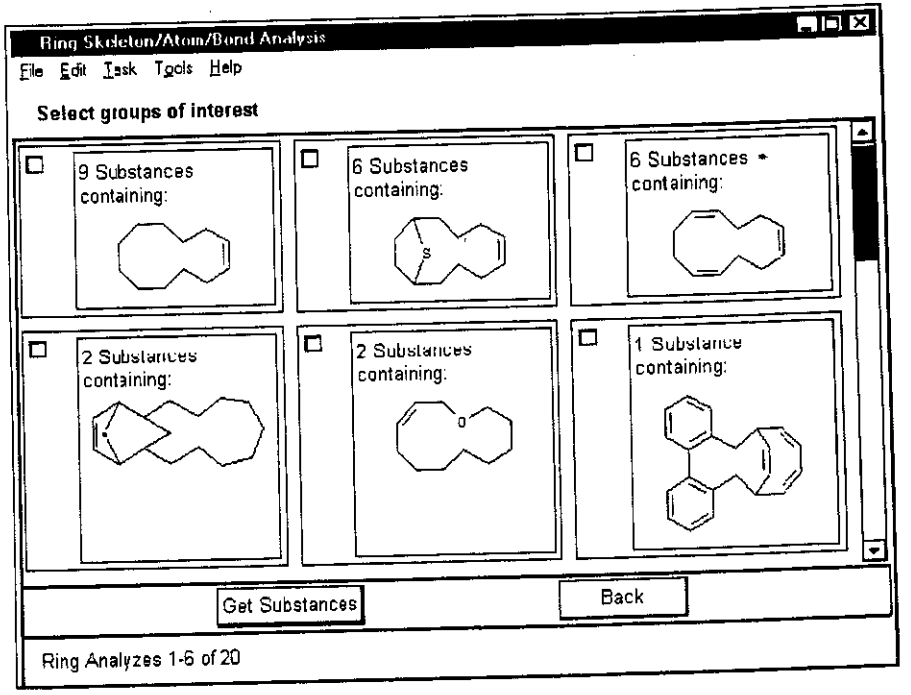



**Figure 6.14** 12-Membered ring skeletons found in search (Figure 6.13). SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

In this example, reactions are first searched in CASREACT and then are refined using reaction refinement tools. Bibliographic records are then obtained in CAPLUS and are analyzed and refined. Finally, SciFinder Scholar and SciFinder subscription users may take the indexed substances back to REGISTRY and so use any of the analyze or refine tools available in the database. The ability to combine structures, reactions, functional groups, and keywords in this way is unique.

Nevertheless, it still helps if the searcher considers alternative approaches! What would happen if the search is commenced in REGISTRY, and how about starting in CAPLUS? The former option would require searches on various cycloalkenes, but general searches would exceed system limits and so this approach would work only if relatively specific structure types are investigated. Meanwhile, to take advantage of the much wider coverage of CAPLUS, an **Explore by Research Topic** of the type: 'I am interested in ''metathesis or Grubbs or RCM with membered'' ' may be considered and indeed, any of the first three concepts are 'closely associated' with the last

concept in almost 200 records, and these could be explored further. SciFinder Scholar and SciFinder subscription users could also consider the option of retrieving any one of the candidate sets from the search in CAPLUS and converting the indexed substances to REGISTRY records, which could be analyzed or refined in the normal way.

## 6.6 Exercises

6.1 Use **Explore by Research Topic** to search for information on the following chemical reactions:

6.1.1 catalysts for asymmetric Heck reaction;

6.1.2 synthesis (preparation) of dihydropyridines by the Hantzsch process;

6.1.3 metal removal (demetalation) of porphyrins;

6.1.4 reaction of propiolates with triethylamine (trimethylamine, tertiary amines);

6.1.5 synthesis of bromotropolones;

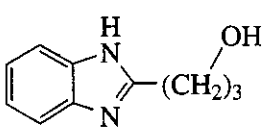
6.1.6 reviews on preparations of transition metal carbonyls;

6.1.7 combinatorial libraries (combinatorial synthesis) of thiazoles;

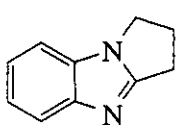
6.1.8 mechanism of ipso nitration of naphthalenes; and

6.1.9 preparations of cycloalkenes by the metathesis reaction (e.g. using Grubbs catalyst).

6.2 The user wants to substitute the OH in (A) with another nucleophile. However, in the laboratory, it is found that any activation of the hydroxyl gives cyclic products of type (B), so protection of the nitrogen is required. How would you use SciFinder Scholar to help with this problem (of protecting a nitrogen, perhaps of an imidazole, in the presence of a primary hydroxyl)?




(A)

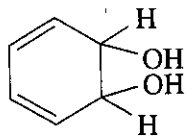



(B)



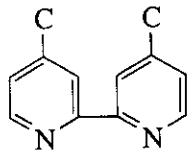
6.3 Explore alternative ways to search for the reaction shown in Figure 6.11 in the reaction database. (Hint: examine possibilities with functional group searches: alkynes (reactant/reagent) and tertiary amines (reactant/reagent). Then, consider **Refine Reactions**, perhaps with part structure C≡C·C(O)O).

6.4 Find reaction information on hydroxylated aromatics of the type (C). (Note: even in CASREACT, this structure will be too general, but try locking *just* the chains bonds.)



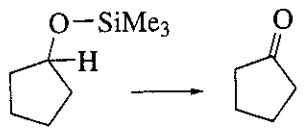

(C)

6.5 Explore retrosynthetic pathways to compounds with the substructure (D).




(D)

6.6 Find references on methods for the one-pot conversion of silyl-protected alcohols to ketones, e.g.




6.7 Find records for the preparations of carboxylic acids by the Pinner reaction.

# Appendices

## Appendix 1 Some SciFinder and SciFinder Scholar Resources

**Resources Published by CAS**

Links to SciFinder and SciFinder Scholar resources may be found through the 'Internet' icon at the top of many screens. That is, extensive information on SciFinder Scholar may be found through the principal CAS Web site: *http://www.cas.org/SCIFINDER/SCHOLAR/index.html*. In particular, the reader is referred to the link to *SciFinder Scholar Resources* from which many links to specific features and solutions may be obtained. The corresponding Web site for *SciFinder* is *http://www.cas.org/SCIFINDER* and resources may be found through *http://www.cas.org/SCIFINDER/CENTRAL /resources.html*.

SciFinder and SciFinder Scholar manuals, which describe the operations of the various functions, may be directly obtained from CAS.

At the top of many of the screens is the 'Help' icon, and from here, numerous help messages may be read.

**ListServe**

Two principal resources are:

- *ScholarTalk@cas.org*
- *CHMINF-L@LISTSERV.INDIANA.EDU*

Library staff interested in contributing to discussions on SciFinder Scholar should definitely participate in these ListServes!

### Information and Training Materials

Many university libraries publish excellent materials on SciFinder Scholar, and a number of these may be found through the ListServes above. Alternatively, searches on the term SCIFINDER through Google, or indeed through SciFinder Scholar, produce numerous materials. Some examples are:

*http://www.leeds.ac.uk/library/documents/workbook/scifinder*
*http://www.lib.rochester.edu/car/scilist.htm*
*http://www.indiana.edu/~cheminfo/cciim31.html*
*http://www.library.ucsb.edu/istl/00-fall/databases.html*
*http://www.chem.ox.ac.uk/cheminfo/WoS_CAS.html*
*http://lamar.colostate.edu/~mwilde/CA/scifinder.html*

As mentioned repeatedly throughout this text, SciFinder Scholar is a comprehensive information resource for nearly all scientific disciplines. Results from research on the coverage of Scholar and comparisons with other resources may be obtained directly by E-mail through: *d.ridley@chem.usyd.edu.au.*

## Appendix 2 CAS Roles in CAPLUS (including changes made in December 2001)

| | |
|---|---|
| **ANST** | **Analytical Study** |
| ANT | Analyte |
| AMX | Analytical Matrix |
| ARG | Analytical Reagent Use |
| ARU | Analytical Role, Unclassified |

| | |
|---|---|
| **BIOL** | **Biological Study** |
| ADV | Adverse Effect, Including Toxicity |
| AGR | Agricultural Use |
| BCP | Biochemical Process[3] |
| BMF | Bioindustrial Manufacture |
| BPN | Biosynthetic Preparation |
| COS | Cosmetic Use[3] |
| DGN | Diagnostic Use[3] |
| DMA | Drug Mechanism of Action[3] |
| FFD | Food or Feed Use |
| NPO | Natural Product Occurrence[3] |
| PAC | Pharmacological Activity[3] |
| PKT | Pharmacokinetics[3] |
| THU | Therapeutic Use |
| BUU | Biological Use, Unclassified |
| BSU | Biological Study, Unclassified**** |

| | |
|---|---|
| **CMBI** | **Combinatorial Study**[1] |
| CPN | Combinatorial Preparation[1] |
| CRT | Combinatorial Reactant[1]** |
| CRG | Combinatorial Reagent[1]*** |
| CUS | Combinatorial Use[1] |
| CST | Combinatorial Study, Unclassified[1] |

| | |
|---|---|
| **FORM** | **Formation, Nonpreparative** |
| GFM | Geological or Astronomical Formation |
| FMU | Formation, Unclassified |

| | |
|---|---|
| **OCCU** | **Occurrence** |
| GOC | Geological or Astronomical Occurrence |
| NPO | Natural Product Occurrence[3] |
| POL | Pollutant |
| OCU | Occurrence, Unclassified |

| | |
|---|---|
| **PREP** | **Preparation** |
| BMF | Bioindustrial Manufacture |
| BPN | Biosynthetic Preparation |
| BYP | Byproduct |
| CPN | Combinatorial Preparation[1] |
| IMF | Industrial Manufacture |
| PUR | Purification or Recovery |
| SPN | Synthetic Preparation |
| PNU | Preparation, Unclassified |

| | |
|---|---|
| **PROC** | **Process** |
| PEP | Physical, Engineering, or Chemical Process |
| CPS | Chemical Process[2]* |
| EPR | Engineering Process[2]* |
| PYP | Physical Process[2]* |
| BCP | Biochemical Process[3] |



| | |
|---|---|
| **PROC** | **Process** |
| GPR | Geological or Astronomical Process |
| REM | Removal or Disposal |

| | |
|---|---|
| **USES** | **Uses** |
| AGR | Agricultural Use |
| ARG | Analytical Reagent Use |
| CAT | Catalyst Use |
| COS | Cosmetic Use[3] |
| CUS | Combinatorial Use[1] |
| DEV | Device Component Use |
| DGN | Diagnostic Use[3] |
| FFD | Food or Feed Use |
| MOA | Modifier or Additive Use |
| POF | Polymer in Formulation |
| TEM | Technical or Engineered Material Use |
| THU | Therapeutic Use |
| BUU | Biological Use, Unclassified |
| NUU | Other Use, Unclassified[4] |

| | |
|---|---|
| **RACT** | **Reactant or Reagent**[2] |
| RCT | Reactant[2] |
| CRT | Combinatorial Reactant[1]** |
| RGT | Reagent[2] |
| CRG | Combinatorial Reagent[1]*** |

**Specific roles not upposted to any super roles**

| | |
|---|---|
| **PRP** | **Properties** |
| **MSC** | **Miscellaneous** |

[1] New role
[2] Split
[3] New in Bio
[4] Name change

* Also has PEP role assigned
** Also has RCT role assigned
*** Also has RGT role assigned
**** For backfile conversion, BSU is assigned *in addition* to former roles BAC, BOC, BPR, MFM

**SciFinder Scholar Note:**

1. Many of these roles are applied when specific references from REGISTRY answer sets are obtained (Figure 3.21).

# Appendix 3 Some Basic Principles used by SciFinder Scholar in the Interpretation of the Research Topic Query

| What Scholar does | What the implications are | Examples |
|---|---|---|
| 1. CONCEPTS | | |
| Scholar uses the prepositions and conjunctions in the question to determine the separate concepts | Generally, it is better to use prepositions and conjunctions between each of the terms in which less than four terms are entered; it usually is not necessary to enter too many terms | Try 'I am interested in ''measurement *of* mass *of* quarks'' ' rather than 'I am interested in ''measurement mass quark'' in the first Explore since use of subsequent refinement tools to narrow answers is preferable |
| If words in the query are not separated by prepositions or conjunctions or certain identifying words, then Scholar identifies the words as a 'single' concept in which answers contain the words 'closely associated' | Searches on exact phrases almost invariably miss important records, and 'closely associated' is a better option; however, the *individual words* in multiword single concepts will not be interpreted also at the broader (whole record) level, which at times may be more appropriate | The entry 'I am interested in ''treatment of wastewater from gold mining'' ' always keeps the terms 'gold mining' in the same sentence |
| Scholar automatically applies truncation and singulars/plurals | This saves considerably on the number of terms needing to be entered, | Research Topic query 'I am interested in ''tropical plants'' ' |

(*continued*)



| What Scholar does | What the implications are | Examples |
|---|---|---|
| | although (quite) rarely the automatic truncation applied may lead to inappropriate hits; a solution is to select relevant records manually | produces hits on 'tropanes from plants', and this may be inappropriate |
| Scholar automatically applies synonyms (including common abbreviations and CAPLUS/MEDLINE abbreviations) from its synonym dictionary | This feature saves considerably on effort required to set up the query and gives more comprehensive results; at times, possibly undesired synonyms may be retrieved and a solution is to select relevant records manually | Research Topic query 'I am interested in ''sheep'' ' produces hits on lamb(s), ram(s), etc. and also on goat(s) |
| 2. CANDIDATES | | |
| Scholar presents a list of candidates and first indicates numbers of answers where *all the concepts* are 'closely associated' (usually in the same sentence) and 'anywhere in the reference' | The concepts identified by Scholar are presented in bold and in quotes, and first a check should be made to see that the concepts are as intended; candidates labeled 'closely associated' and 'anywhere in the reference' simply give users alternatives relating to the closeness of the terms, the assumption being that the closer the terms, the more directly they are related | The entry 'I am interested in ''liver or kidney diseases'' ' identifies concepts **'liver'** and **'kidney diseases'**, which is not the intention (the issue is addressed below, but the point here is that *it helps to check the concepts!*) |
| Scholar next indicates number of answers in which combinations of *some of the* | If three concepts (A, B, C) are identified, then the number of records for candidates with | The entry 'I am interested in ''the reaction of propiolates with |

(*continued*)

| What Scholar does | What the implications are | Examples |
|---|---|---|
| *concepts* are 'closely associated' or 'anywhere in the reference' | combinations A and B, A and C, and B and C will be listed; often, fewer concepts will give better results simply because *all of the concepts may not be* in the record | amines'' ' gives around 20 hits where all three concepts are 'closely associated', but almost double the hits for just 'propiolates' and 'amines' 'closely associated'. May this be a better option? |
| Scholar next indicates number of answer candidates *for the individual concepts* | The greater the number of concepts, the greater the restrictions on the answers, and it is most helpful to see listings for the individual concepts, particularly when few records are identified with all the concepts present | The entry 'I am interested in ''wastewater from gold mine tailings'' ' shows very few hits for the individual separate concept 'gold mine tailings' . . . and the user immediately identifies where the potential difficulty is |
| 3. SYNONYMS | | |
| Users may force inclusion of alternative terms by adding the terms in parentheses | Care must be taken with 'distributed modifiers' (below). Parentheses are ignored under Refine options (AND operator applies to terms). | Use the Research Topic query 'I am interested in ''chiral reduction (chiral hydrogenation)'' ' rather than 'I am interested in chiral reduction (hydrogenation)' |



| What Scholar does | What the implications are | Examples |
|---|---|---|
| 4. BOOLEAN OPERATORS | | |
| Scholar interprets the Boolean AND as a request for both terms *anywhere* in the reference, and the 'closely associated' option is not presented | Use of AND rather than a preposition will not alert the user to the 'closely associated' option | It is better to use entry 'I am interested in ''mass of quarks'' ' rather than 'I am interested in ''mass and quarks'' ' |
| In some instances, Scholar may interpret AND as OR | However, it is safer to use OR when it is meant! | Research Topic query 'I am interested in ''sugars and carbohydrates'' ' will give candidates for OR as well as AND |
| Scholar interprets the Boolean OR to signify either terms | When entering terms under **Explore by Research Topic**, it does not matter whether alternative terms are placed in parentheses or linked with OR. However, parentheses are ignored under Refine options (AND operator applies to terms). | The entries 'I am interested in ''steroid analysis in urine of men or women or humans'' ' or 'I am interested in ''steroid analysis in urine of men (women, humans)'' ' give similar results |
| Scholar interprets the Boolean NOT to exclude records with the terms entered | The query is read literally from left to right, so answer sets depend on the placement of the operator | The entry 'I am interested in ''tea and sugar not coffee'' is' interpreted differently from 'I am interested in ''tea not coffee and sugar'' ' |

| What Scholar does | What the implications are | Examples |
|---|---|---|
| Scholar does not distribute modifiers | Modifiers (usually adjectives) need to be entered with each term to which they refer | The entry 'I am interested in ''chiral reduction (hydrogenation)'' ' identifies concepts '**chiral reduction**' and '**hydrogenation**', so 'I am interested in ''chiral reduction (chiral hydrogenation)'' ' should be used |

# Appendix 4 Examples of Analyze References

Some further examples of **Analyze References** (see Chapter 2, Section 2.5.4). Data is from answer set in CAPLUS for search 'I am interested in ''inhibition of HIV replication in humans'' '.

**Analyze References** by **CAS Registry Number** gives histograms that indicate key substances. It differs from **Get Related** then **Substances,** which gives REGISTRY answers that may be analyzed or refined.

**Analyze References** by **CAS Section Title** is an excellent way to narrow answers by *broad categories*.

**Analyze References** by **Document Type** allows scientists to review options. Note here that about 16 % of the records are from patents. MEDLINE has more than 40 document types including clinical trial (>310 000 records), commentary (>180 000), letter (>440 000), and review of reported cases (>42 000).

**Analyze References** by **Journal Name** is valuable for the identification of key journals.

**Analyze References** by **Language**. Language is that of the original document.

**Analyze References** by **Publication Year** is an excellent way to review research trends and to narrow answer sets into conveniently small groups. Analyze 'alphabetically' gives histograms in reverse chronological order.



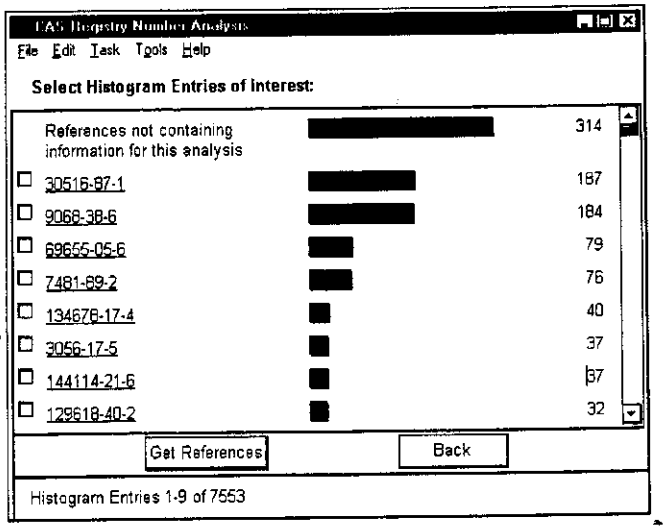



**Figure A4.1** Screen for **CAS Registry Number Analysis**. SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

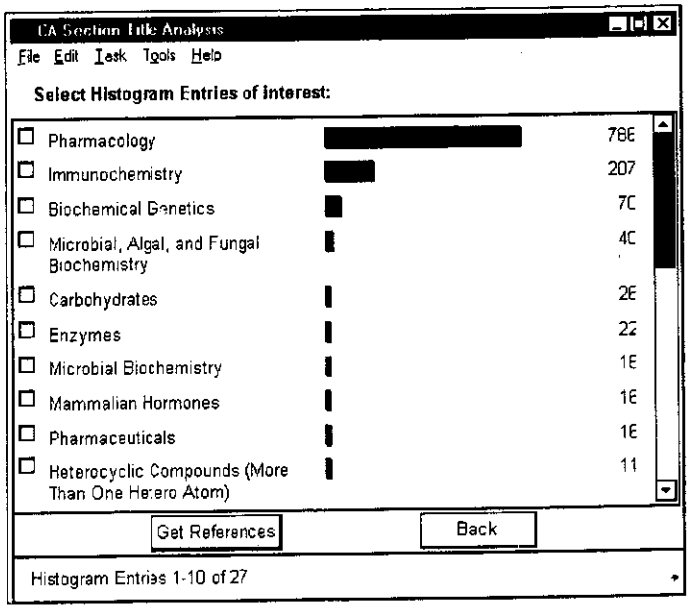



**Figure A4.2** Screen for **CA Section Title Analysis**. SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

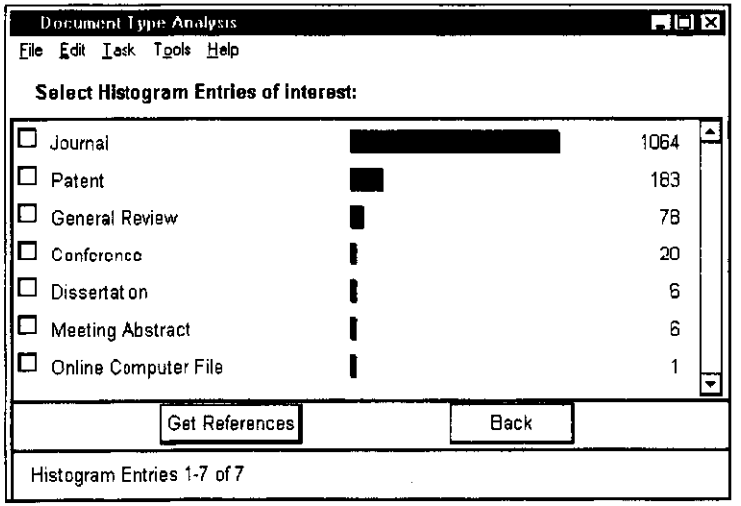



**Figure A4.3** Screen for **Document Type Analysis**. SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

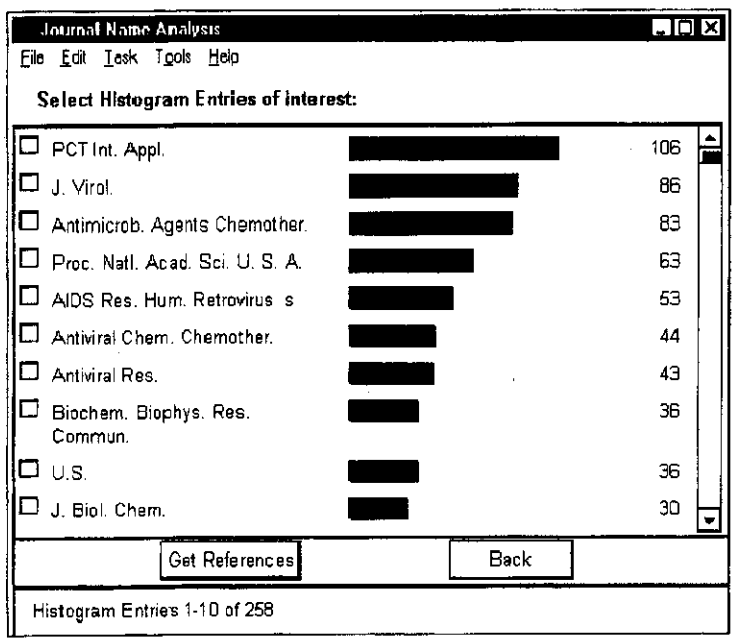



**Figure A4.4** Screen for **Journal Name Analysis**. SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.



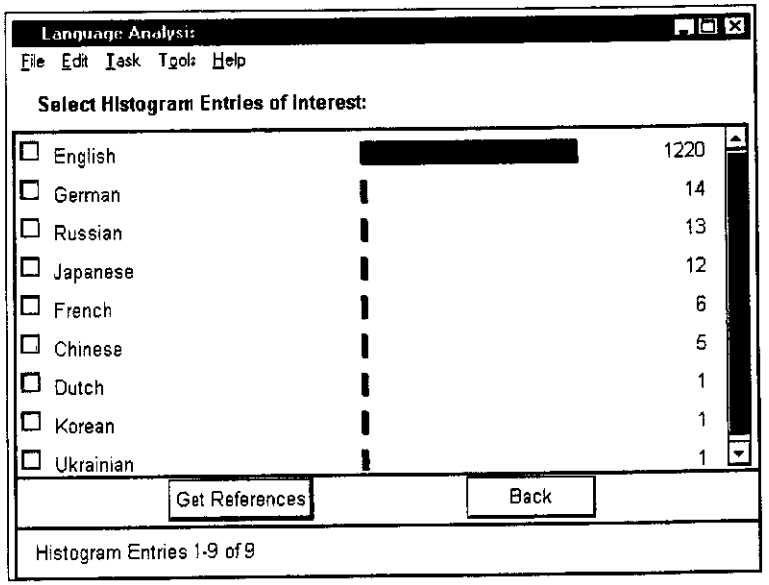



**Figure A4.5** Screen for **Language Analysis**. SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

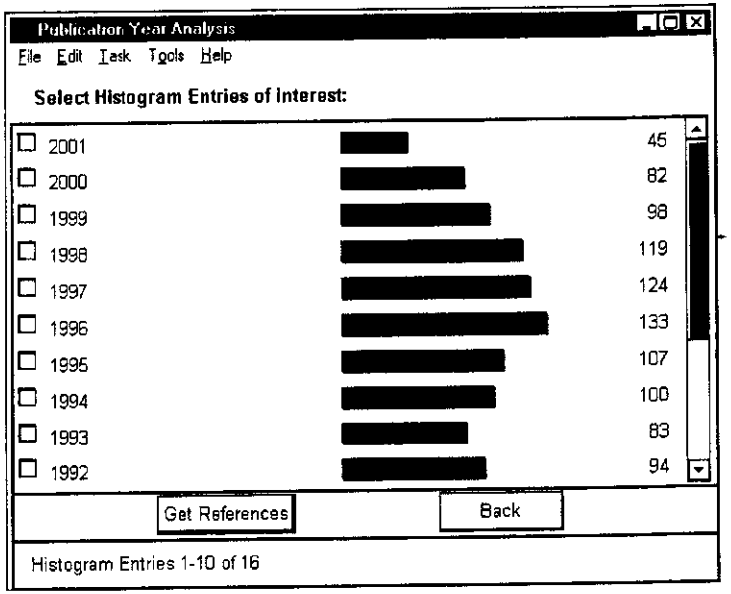



**Figure A4.6** Screen for **Publication Year Analysis**. SciFinder Scholar screens are reproduced with permission of the Chemical Abstracts Service, a division of the American Chemical Society.

# Appendix 5

**Registration of Substances**

The substance database searched in SciFinder Scholar is REGISTRY. It includes all types of chemical substances (organic, inorganic, alloys, polymers, mixtures, reactive intermediates, proteins, nucleic acids, etc.). In most cases, the registration of substances follows exactly the valence bond descriptions taught in chemistry courses. However, the many subtle variations in structures often require modification of valence bond descriptions and the key issues are resonance, tautomerism, $\pi$-complexes, $\sigma$-complexes, radicals, and other reactive intermediates. Specific rules need to be applied to handle these modifications when structures are input into computer databases.

Computer databases also need to have rules for defining, *inter alia*, salts, mixtures, hydrates, polymers, and alloys in which valence bond descriptions do not necessarily apply. Many of these are addressed by defining a single registered substance as one made up of a variety of components. About 10% of the registrations are multicomponent substances, and the key is to recognize that each component is identified as a single entry in the formula and in the structure fields.

SciFinder Scholar automatically interprets all these situations and provides answers, although users not familiar with the computer conventions may initially be confused by the types of answers produced. A good way to learn about these issues is to examine actual records and this appendix gives examples. Except for Appendix 5.1.1, which gives the complete record, only key parts of the records are given.

## A5.1 Single-Component Substances

### *A5.1.1 Single Substances*

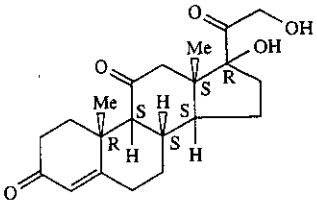
The most common class is that of single substances, and a typical record for a single substance is shown. The CAS systematic names, common usage names, and trade names are listed first. Then, most records list files on the STN network *http://www.cas.org/stn.html*, which give further information on the substance. The files automatically included in SciFinder Scholar are CAPLUS (the CAS bibliographic database), CASREACT (the chemical reaction database), CHEMCATS (commercial substances database), CHEMLIST (regulated chemicals database), and MEDLINE (the NLM bibliographic database).



**Registry Number:** 53-06-5
**CA Index Name:** Pregn-4-ene-3,11,20-trione, 17,21-dihydroxy- (7CI, 9CI)
**Other Names:** Cortisone (8CI); 11-Dehydro-17-hydroxycorticosterone; 17-Hydroxy-11-dehydrocorticosterone; 17$\alpha$-Hydroxy-11-dehydrocorticosterone; Adrenalex; Compound E; Cortisate; Cortivite; Cortogen; Cortone; KE; Kendall's compound E; Reichstein's substance Fa; Wintersteiner's compound F
**Formula:** C21 H28 O5
**STN Files:** CAPLUS, AGRICOLA, AIDSLINE, ANABSTR, BEILSTEIN, BIOBUSINESS, BIOSIS, BIOTECHNO, CA, CABA, CANCERLIT, CAOLD, CASREACT, CBNB, CEN, CHEMCATS, CHEMINFORMRX, CHEMLIST, CIN, CSCHEM, CSNB, DDFU, DETHERM, DRUGU, EMBASE, HODOC, IFICDB, IFIPAT, IFIUDB, IPA, MEDLINE, MRCK, MSDS-OHS, NAPRNOTE:, NIOSHTIC, PIRA, PROMT, RTECS, SPECINFO, TOXLINE, TOXLIT, USAN, USPATFULL, VETU

(Additional Information is available through STN International. Contact your information specialist, a local CAS representative, or the CAS Help Desk for Assistance)

Absolute stereochemistry.




**~3253 References**
**Database:** REGISTRY

***SciFinder Scholar Note:***

1. In a search on a name through **Explore by Chemical Substance** then **Substance Identifier**, SciFinder Scholar initially seeks a match with the CA Index Name or any of the Other Names. If there is not an exact match, then an attempt will be made to find possible answers on the basis of parts of names (Chapter 3, Section 3.6).

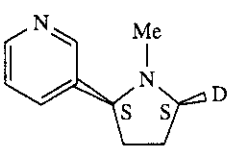
2. When terms are entered under **Explore by Research Topic**, SciFinder Scholar checks to see whether any of the terms exactly matches the CA Index Name or any of the Other Names. If there is an exact match, the CAS Registry Number is included in the 'concept' searched.

## *A5.1.2 Isotopic Substances*

Isotopes of hydrogen are indicated by the symbols D and T, while isotopes of other atoms have the atomic weight as a superscript before the symbol for the atom. These representations appear in the structure diagram, in the name entries, and, in the cases of the hydrogen isotopes, in the formula.

**Isotopes of Hydrogen**

| | |
|---|---|
| **Registry Number:** | 105089-96-1 |
| **CA Index Name:** | Pyridine, 3-(1-methyl-2-pyrrolidinyl-5-d)-, (2*S*-*cis*)- (9CI) |
| **Formula:** | C10 H13 D N2 |
| | Absolute stereochemistry. |

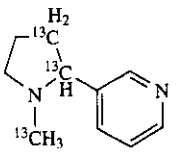







**Isotopes of other Elements (here carbon)**

| | |
|---|---|
| **Registry Number:** | 67209-85-2 |
| **CA Index Name:** | Pyridine, 3-[1-(methyl-13C)-2-pyrrolidinyl-2,3-13C2]- (9CI) |
| **Other Names:** | Pyridine, 3-[1-(methyl-13C)-2-pyrrolidinyl-2,3-13C2]-, ... |
| **Formula:** | C10 H14 N2 |






***SciFinder Scholar Note:***

1. Isotopes of hydrogen (D, T) are included in the molecular formula, so **Explore by Chemical Substance** then **Molecular Formula** may be used.

2. Other isotopic substances are best found by searching on the structure (exact/related) and then scrolling through the answers.

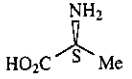
## *A5.1.3 Stereoisomers*

The letters R, S, E, and Z, and the Greek letters $\alpha$ and $\beta$ are commonly used to indicate stereochemical arrangements in molecules. Structure diagrams additionally indicate stereochemistry through heavy, dotted, or wedged bonds.

An example of the registration of an enantiomer and of a racemic form are shown. In particular, the stereochemical descriptors should be noted in the name fields, namely, the (S), L, and (+) descriptors and the (R,S), DL, dl, and (.+−.) descriptors.

**Single Stereoisomer**

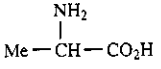
| | |
|---|---|
| **Registry Number:** | 56-41-7 |
| **CA Index Name:** | L-Alanine (9CI) |
| **Other Names:** | Alanine, L- (7CI, 8CI); (*S*)-(+)-Alanine; (*S*)-2-Aminopropanoic acid; (*S*)-Alanine; α-Alanine; α-Aminopropionic acid; Alanine; L-(+)-Alanine; L-α-Alanine; L-α-Aminopropionic acid; L-2-Aminopropanoic acid; L-2-Aminopropionic acid; Propanoic acid, 2-amino-, (*S*)- |
| **Formula:** | C3 H7 N O2 |
| | Absolute stereochemistry. Rotation (+). |






**Racemate**

| | |
|---|---|
| **Registry Number:** | 302-72-7 |
| **CA Index Name:** | Alanine (9CI) |
| **Other Names:** | Alanine, DL- (8CI); DL-Alanine; (±)-2-Aminopropionic acid; (±)-Alanine; (*R*,*S*)-Alanine; DL-α-Alanine; DL-α-Aminopropionic acid; dl-2-Aminopropanoic acid; dl-Alanine |
| **Formula:** | C3 H7 N O2 |








Meanwhile, cortisone (Appendix 5.1.1 augmentin (Appendix 5.2.3), and nicotine hydroiodide (Appendix 5.2.1) have stereochemical features, and the ways in which these are presented in the name and structure fields should be noted.

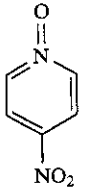
***SciFinder Scholar Note:***

1. Specific stereoisomers may be searched throught structure searches (see update through *http://www.wiley.co.uk/ridley*).

## *A5.1.4 Donor Bonds*

When one of the atoms in the bond provides both the bond electrons, the structure is represented with a double bond. This occurs in particular with higher oxidation states of metals and of nonmetals such as nitrogen, phosphorus, and sulfur. In some instances, this representation will produce a 'structure' that violates valence bond rules (here, the valence of 4 for nitrogen).

| | |
|---|---|
| **Registry Number:** | 1124-33-0 |
| **CA Index Name:** | Pyridine, 4-nitro-, 1-oxide (6CI, 8CI, 9CI) |
| **Other Names:** | 4-Nitropyridine 1-oxide; 4-Nitropyridine N-oxide; 4-Nitropyridine oxide; *p*-Nitropyridine N-oxide |
| **Formula:** | C5 H4 N2 O3 |






***SciFinder Scholar Note:***

1. SciFinder Scholar automatically recognizes compounds with donor bonds and searches the appropriate structure. For example, the substance above is retrieved irrespective of whether a double or a single bond is drawn between the nitrogen and the oxygen.

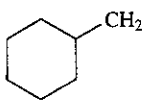
### *A5.1.5 Intermediates*

Examples of a carbene and a radical are shown. However, the CAS Registry Number for reactive intermediates will be inserted in CAPLUS only when the intermediates are clearly identified or are an important part of the original paper. Intermediates simply drawn in mechanistic schemes are not indexed.

**Registry Number:** 39922-09-3
**CA Index Name:** Cyclohexylidene (9CI)
**Other Names:** Carbenacyclohexane
**Formula:** C6 H10




**Registry Number:** 67271-34-5
**CA Index Name:** Methyl, cyclohexyl- (9CI)
**Other Names:** Cyclohexylmethyl radical
**Formula:** C7 H13






***SciFinder Scholar Note:***

1. Radicals, carbocations, and carbanions have relatively unusual formulas (because of the odd valency at carbon), so **Explore by Chemical Substance** then **Molecular Formula** usually retrieves the intermediates quickly. If needed, **Refine Substances** followed by drawing, then searching, on the carbon skeleton may further narrow answer sets.
2. Carbenes are isomeric with alkenes, so initial searches on molecular formulas will produce larger answer sets that need to be refined (e.g. by structure).



3. Because of the odd valency of carbon in these intermediates, substructure searches on the ring skeletons may not retrieve intermediates, so an initial molecular formula search may be necessary.

## A5.2 Multicomponent Substances

A record for a multicomponent substance lists only a few of the names for the individual substances involved, has a 'dot disconnected' entry in the formula field, and has the individual components presented in the structure display (which also gives the CAS Registry Number of the component). Multicomponent substances are commonly encountered in polymers that have greater than one monomer constituent, in salts, mixtures, minerals, and alloys.

At present, in the substance database, there are around 3.8 million substances with two or more components.

### *A5.2.1 Salts*

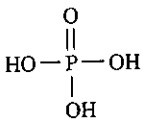
**Salts that do not Contain Carbon**

Simple salts from acids not containing periodic table group VI atoms (oxygen, sulfur, etc.) are registered in the way normally drawn by the chemist. For example, sodium chloride and calcium bromide are represented as NaCl and $CaBr_2$, respectively.

However, salts from acids containing periodic table group VI atoms and bases containing periodic table group I and II atoms (sodium, potassium, calcium, barium, etc.) are registered as the free acid combined with the base (which is given the metal symbol). For example, although chemists write the formula for the calcium phosphate as $Ca_3(PO_4)_2$, the substance is registered as if the hydrogens were still attached to the phosphate group, that is, as phosphoric acid, $H_3PO_4$. As the actual salt has three Ca atoms to two phosphate groups, the substance is considered as $Ca_3(H_3PO_4)_2$ and the molecular formula, which has the first component in the alphabet as a single atom, then becomes Ca.2/3 H3O4P.

Actually, there are two substances called *calcium phosphate* in the database and only one is shown here (the Molecular Formula Field entry for the second substance is: MF Ca. x H3 O4 P.) The issue is the chemical description of the substances in the original literature. If the particular form is not specified in the article, the indexer will register the substance as that in which the ratio is not specified. (There are currently more than 2000 entries for this second substance, so a search for information on calcium phosphate should probably include its Registry Number (10103-46-5) as well!)

| | |
|---|---|
| **Registry Number:** | 7758-87-4 |
| **CA Index Name:** | Phosphoric acid, calcium salt (2:3) (8CI, 9CI) |
| **Other Names:** | . . . Calcium phosphate; Calcium phosphate (3:2); Calcium phosphate (Ca3(PO4)2); Calcium tertiary phosphate; Tribasic calcium phosphate; Tricalcium diphosphate; Tricalcium orthophosphate; Tricalcium phosphate; Tricalcium phosphate (Ca3(PO4)2) . . . |
| **Formula:** | Ca. 2/3 H3 O4 P |



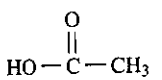
· 3/2 Ca



**Salts from Organic Acids and Periodic Table Group I/II Bases**

The registration is similar to the registration described above. Note the molecular formula for sodium acetate is listed as C2H4O.Na, whereas a chemist would represent the formula for the substance as $C_2H_3ONa$.

| | |
|---|---|
| **Registry Number:** | 127-09-3 |
| **CA Index Name:** | Acetic acid, sodium salt (7CI, 8CI, 9CI) |
| **Other Names:** | Sodium acetate (6CI); Anhydrous sodium acetate |
| **Formula:** | C2 H4 O2. Na |



· Na

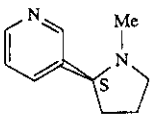




**Salts with Nitrogen-Containing Bases**

Apart from those involving periodic table group I/II atoms, the most common bases are those involving nitrogen atoms, and the salts are represented as the free base and the free acid (which includes the hydrogen donated by the acid).

| | |
|---|---|
| **Registry Number:** | 6019-03-0 |
| **CA Index Name:** | Pyridine, 3-(1-methyl-2-pyrrolidinyl)-, monohydriodide, (S)- (9CI) |
| **Other Names:** | Nicotine, monohydriodide (8CI); Nicotine hydriodide; Nicotine hydroiodide |
| **Formula:** | C10 H14 N2. H I |
| | Absolute stereochemistry. |



· HI



Salts from nitrogen-containing bases involving other acids (e.g. sulfuric acid and phosphoric acid) are registered similarly.

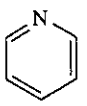
**Salts from Organic Acids and Organic Bases**

These salts are registered as the free acid and the free base, but the subtle difference is that now both the acid and the base are registered as individual components.

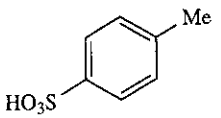
| | |
|---|---|
| **Registry Number:** | 24057-28-1 |
| **CA Index Name:** | Benzenesulfonic acid, 4-methyl-, compd. with pyridine (1:1) (9CI) |
| **Other Names:** | Pyridine, 4-methylbenzenesulfonate; Pyridine, *p*-toluenesulfonate (7CI, 8CI); PPTS; Pyridinium 4-toluenesulfonate; Pyridinium *p*-toluenesulfonate; |

(*continued*)

| | |
|---|---|
| | Pyridinium *p*-tosylate; Pyridinium tosylate |
| **Formula:** | C7 H8 O3 S. C5 H5 N |
| Component Registry Number: | 110-86-1 |
| Formula: | C5 H5 N |




| | |
|---|---|
| Component Registry Number: | 104-15-4 |
| Formula: | C7 H8 O3 S |






***SciFinder Scholar Note:***

1. Many biologically important substances are salts and have different CAS Registry numbers from the parent acid or base, so it is important to consider search strategies that will retrieve the parent substance including all its salts.
2. Salts are retrieved through an exact/related structure search on either the parent acid or base.
3. To retrieve specific salts, draw both structure fragments on the same screen and then perform an exact/related structure search.

### *A5.2.2 Alloys*

When specified in the original article, the composition of the alloy is listed in the name and composition fields and the constituent elements are listed in the formula field. A more generic description is applied when the elemental composition is not precisely specified in the original article.



| | |
|---|---|
| **Registry Number:** | 11109-50-5 |
| **CA Index Name:** | Iron alloy, base, Fe 66-74,Cr 18.00-20.00,Ni 8.00-10.50,Mn 0-2.00,Si 0-1.00,C 0-0.08,P 0-0.045,S 0-0.030 (UNS S30400) (9CI) |
| **Other Names:** | (Include) . . . Nickel alloy, nonbase, 1809; CSN 17240; DIN 1.2781 . . . |
| **Formula:** | C. Cr. Fe. Mn. Ni. P. S. Si |
| **Class Identifier:** | Alloy |

**Composition:**

| Component Percent | Component | Component Registry Number |
|---|---|---|
| Fe | 66–74 | 7439-89-6 |
| Cr | 18.00–20.00 | 7440-47-3 |
| Ni | 8.00–10.50 | 7440-02-0 |
| Mn | 0–2.00 | 7439-96-5 |
| Si | 0–1.00 | 7440-21-3 |
| C | 0–0.08 | 7440-44-0 |
| P | 0–0.045 | 7723-14-0 |
| S | 0–0.030 | 7704-34-9 |



| | |
|---|---|
| **Registry Number:** | 12597-68-1 |
| **CA Index Name:** | Stainless steel (9CI) |
| **Other Names:** | Austenitic stainless steel; Chromium nickel stainless steel; Chromium stainless steel; Nickel stainless steel; Pall Ring; STN 350 |
| **Formula:** | Unspecified |
| **Class Identifier:** | Alloy, Manual Registration |
| | No Structure Diagram Available |

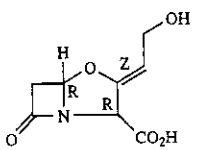
***SciFinder Scholar Note:***

1. Most alloys are easily retrieved through molecular formula searches, and if special compositions are required, it is necessary to look through individual records.
2. Alloys may also be retrieved by drawing the separate elements in the structure screen and then exploring exact/related or substructure search options as needed.

### *A5.2.3 Mixtures*

Mixtures are registered where two or more chemically discrete components have been mixed together for a specific use (e.g. with pharmaceutical and agricultural chemicals). In general, host–guest complexes also are considered as mixtures.

**Registry Number:** 74469-00-4
**CA Index Name:** 4-Oxa-1-azabicyclo[3.2.0]heptane-2-carboxylic acid, 3-(2-hydroxyethylidene)-7-oxo-, monopotassium salt, (2R,3Z,5R)-, mixt. with (2*S*,5*R*,6*R*)-6-[[(2*R*)-amino(4-hydroxyphenyl)acetyl]amino]-3,3-dimethyl-7-oxo-4-thia-1-azabicyclo[3.2.0]heptane-2-carboxylic acid (9CI)
**Other Names:** (*Include*) Amoxicillin-potassium clavulanate mixt.; Augmentin; Augmentin (antibiotic); BRL 25000; BRL 25000A; BRL 25000 G
**Formula:** C16 H19 N3 O5 S. C8 H9 N O5. K
**Class Identifier:** Mixture
Component Registry Number: 61177-45-5 (58001-44-8)
Absolute stereochemistry. Double bond geometry as shown.
Formula: C8 H9 N O5. K

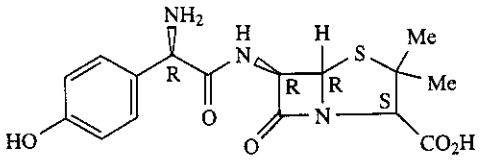



K

(*continued*)



Component Registry Number: 26787-78-0
Absolute stereochemistry.
Formula: C16 H19 N3 O5 S

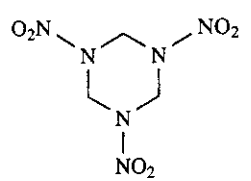





**Registry Number:** 67713-16-0
**CA Index Name:** 1,3,5-Triazine, hexahydro-1,3,5-trinitro-, mixt. with aluminum and 2-methyl-1,3,5-trinitrobenzene (9CI)
**Other Names:** Aluminum, mixt. contg. (9CI); Benzene, 2-methyl-1,3,5-trinitro-, mixt. contg. (9CI); Composition A; Composition B; Composition C; Composition D; H 6; HBX 3; Hexotonal; KS 22; Torpex
**Formula:** C7 H5 N3 O6. C3 H6 N6 O6. Al
**Class Identifier:** Mixture
Component Registry Number: 7429-90-5
Formula: Al

Al

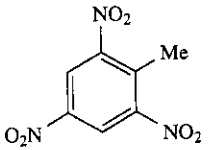
Component Registry Number: 121-82-4
Formula: C3 H6 N6 O6




Component Registry Number: 118-96-7

(*continued*)

Formula: C7 H5 N3 O6






***SciFinder Scholar Note:***

1. An exact/related structure search will retrieve substances in which the substance is a component of a mixture.

2. To search for specific multicomponent substances, draw the separate components on the structure drawing screen and search 'exact/related'.

## A5.3 Metal Complexes

The representation of the structures of some coordination compounds requires modifications to normal valence bond definitions. Generally, the electrons involved in bonding the organic groups to the metal are provided by the organic groups, and in the language of the discipline, there are two types of donors, $\sigma$-donors and $\pi$-donors. In broad terms, these are distinguished in that the electrons in the former case come from atoms, whereas in the latter case, they come from double bonds.

Another issue is that the difference between a coordination compound and a salt may be difficult to define. The following examples are illustrative.

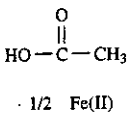
### *A5.3.1 Salts*

**Registry Number:** 3094-87-9
**CA Index Name:** Acetic acid, iron(2+) salt (8CI, 9CI)
**Other Names:** Ferrous acetate; Iron acetate [Fe(OAc)2]; Iron diacetate; Iron(2+) acetate; Iron(II) acetate

(*continued*)



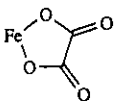
**Formula:** C2 H4 O2. 1/2 Fe






### *A5.3.2 Coordination Compounds*

**Registry Number:** 516-03-0
**CA Index Name:** Iron, [ethanedioato(2-)-□ O1, □ O2]-(9CI)
**Other Names:** Iron, [ethanedioato(2-)-O,O']-; Oxalic acid, iron(2+) salt (1:1) (8CI); Ferrous oxalate; Ferrous oxalate (1:1); Ferrous oxalate (Fe(C2O4)); Ferrox; Iron oxalate; Iron protoxalate; Iron(2+) oxalate; Iron(II) oxalate; Oxalic acid, iron(2+) salt
**Formula:** C2 Fe O4





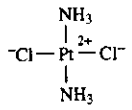

### *A5.3.3 σ-Complexes*

Generally, charges relate to the species involved in the preparation of the complex. For example, cisplatin is made from $Pt^{2+}$, $Cl^-$, and $NH_3$.

**Registry Number:** 15663-27-1
**CA Index Name:** Platinum, diamminedichloro-, (SP-4-2)- (9CI)
**Other Names:** (Include) Platinum, diamminedichloro-, *cis* -(8CI); *cis* -Platinum II; *cis* -Platinum(II) diaminodichloride; *cis* -Platinum(II) diamminedichloride; *cis* -Platinumdiamine dichloride; *cis* -Platinumdiammine dichloride; Cisplatin; Cis-platinum; Cisplatyl; CPDD; CPPD; DDP

(*continued*)

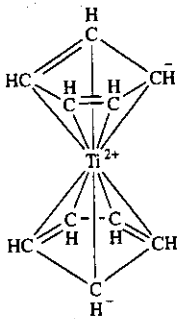
| | |
|---|---|
| **Formula:** | Cl2 H6 N2 Pt |
| **Class Identifier:** | Coordination Compound |





### *A5.3.4 π-Complexes*

| | |
|---|---|
| **Registry Number:** | 1271-29-0 |
| **CA Index Name:** | Titanocene (9CI) |
| **Other Names:** | Titanium, di-π -cyclopentadienyl- (8CI); Titanium, dicyclopentadienyl- (6CI, 7CI); Dicyclopentadienyltitanium ' |
| **Formula:** | C10 H10 Ti |
| **Class Identifier:** | Coordination Compound |





### *SciFinder Scholar Note:*

1. Bonds to metals in queries may be ignored in the initial structure search. In these cases, **Analyze Substances** by **Precision** may be required to obtain more precise answer sets.
2. It is not necessary to insert charges in structure queries.



## A5.4 Macromolecules

There are many ways in which polymers are made and these can often determine how the polymers are registered. For example, homopolymers, copolymers, posttreated polymers, block polymers, graft polymers, and polymer blends have specific indexing rules, and examples of some are given.

### *A5.4.1 Homopolymers*

| | |
|---|---|
| **Registry Number:** | 9003-53-6 |
| **CA Index Name:** | Benzene, ethenyl-, homopolymer (9CI) |
| **Other Names:** | (Include over 1000 names. . .) |
| **Formula:** | (C8 H8)*x* |
| **Class Identifier:** | Polymer |
| **Polymer Class Term:** | Polystyrene |

Component Registry Number: 100-42-5

Formula: C8 H8 $H_2C=CH-Ph$

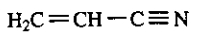


### *A5.4.2 Copolymers*

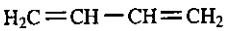
| | |
|---|---|
| **Registry Number:** | 9003-56-9 |
| **CA Index Name:** | 2-Propenenitrile, polymer with 1,3-butadiene and ethenylbenzene (9CI) |
| **Other Names:** | *Over 1,000 names including. . .* 1,3-Butadiene, polymer with ethenylbenzene and 2-propenenitrile (9CI); Acrylonitrile, polymer with 1,3-butadiene and styrene (8CI); Benzene, ethenyl-, polymer with 1,3-butadiene and 2-propenenitrile (9CI); Styrene, polymer with acrylonitrile and 1,3-butadiene (8CI); ABS; ABS (polymer); ABS 1 |
| **Formula:** | (C8 H8. C4 H6. C3 H3 N)x |
| **Class Identifier:** | Polymer |
| **Polymer Class Term:** | Polyacrylic, Polyolefin, Polystyrene |

(*continued*)

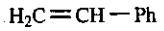
Component Registry Number: 107-13-1
Formula: C3 H3 N



Component Registry Number: 106-99-0
Formula: C4 H6



Component Registry Number: 100-42-5
Formula: C8 H8

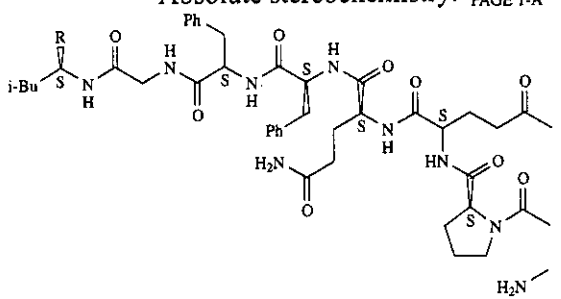




### *A5.4.3 Proteins*

**Registry Number:** 33507-63-0
**CA Index Name:** Substance P (9CI)
**Other Names:** L-Methioninamide, L-arginyl-L-prolyl-L-lysyl-L-prolyl-L-glutaminyl-L-glutaminyl-L-phenylalanyl-L-phenylalanylglycyl-L-leucyl-; Neurokinin P; Substance P (1-11); Substance P (peptide); Substance P (smooth-muscle stimulant)
**Formula:** C63 H98 N18 O13 S

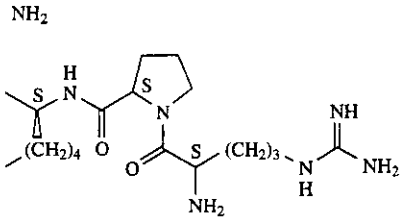
Absolute stereochemistry. PAGE 1-A



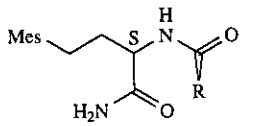
PAGE 1-B

*(continued)*





PAGE 2-A



**Biosequence Length:** 11
**Biosequence Note:** modified

| type | location | description |
|---|---|---|
| terminal mod. | Met-11 | - C-terminal amide |

**Biosequence:**
1 RPKPQQFFGL M



### *A5.4.4 Nucleic Acids and Related Substances*

The CAS register all substances as precisely as possible. Accordingly, if two nucleic acid sequences differ by even one nucleic acid base, then different CAS Registry Numbers are used. This has considerable implication in the rapidly developing area of molecular biology, and searchers should be particularly careful to ensure that all appropriate CAS Registry Numbers are retrieved.

**Registry Number:** 147178-97-0
**CA Index Name:** Guanosine, 2′-deoxycytidylyl-(3′ → 5′)-2′-deoxyguanylyl-(3′ → 5′)-2′-deoxycytidylyl-(3′ → 5′)-2′-deoxyguanylyl-(3′ →

*(continued)*

| | |
|---|---|
| | 5′)-2′-deoxyadenylyl-(3′ → 5′)-2′-deoxyadenylyl-(3′ → 5′)-thymidylyl-(3′ → 5′)-thymidylyl-(3′ → 5′)-2′-deoxycytidylyl-(3′ → 5′)-2′-deoxyguanylyl-(3′ → 5′)-2′-deoxycytidylyl-(3′ → 5′)-2′-deoxy-, double-stranded complementary (9CI) |
| **Other Names:** | GenBank I54951 |
| **Formula:** | Unspecified |
| **Class Identifier:** | Manual Registration |
| | No Structure Diagram Available |
| **Biosequence Length:** | 12 |
| **Biosequence Note:** | double-stranded |
| **Biosequence:** | |
| | 1 cgcgaattcg cg |



### *SciFinder Scholar Note:*

1. Homopolymers may be searched through molecular formulas.
2. Copolymers may be searched through structure queries with the separate components.
3. SciFinder users may employ BLAST® algorithms to retrieve sequences.

## A5.5 Other Cases

### *A5.5.1 Incompletely Defined Substances*

Incompletely defined substances are those that have a known molecular formula but for which the complete valence bond structure was not fully described in the original article. For example, while *o*−, *m*−, and *p*-xylene are the specific dimethylbenzenes, if only 'xylene' is mentioned in the original article, then the incompletely defined substance will be indexed. Similar issues are encountered with salts in which ions have different possible ratios.

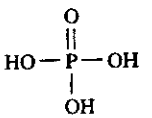


| | |
|---|---|
| **Registry Number:** | 1330-20-7 |
| **CA Index Name:** | Benzene, dimethyl- (9CI) |
| **Other Names:** | Xylene (8CI); Dilan; Dimethylbenzene; Xylol |
| **Formula:** | C8 H10 |
| **Class Identifier:** | Incompletely Defined Substance |

2 (D1—Me)



| | |
|---|---|
| **Registry Number:** | 10103-46-5 |
| **CA Index Name:** | Phosphoric acid, calcium salt (8CI, 9CI) |
| **Other Names:** | 1: PN: WO9961057 SEQID: 2 claimed sequence; Calcium phosphate; Dikal 21; Dynafos; KDV 15u; LF Bosei CP-Z; LF-CP-ZA; Man-Gill 51504 |
| **Formula:** | Ca. *x* H3 O4 P |



· X Ca



### *SciFinder Scholar Note:*

1. Generally, incompletely defined substances will be retrieved in substructure searches when the structure searched is part of the incomplete structure for the substance in the database.

### *A5.5.2 Minerals*

Naturally occurring minerals have many different chemical compositions or crystalline forms and may be registered in a variety of ways. Usually, they are best retrieved through name-based search terms.

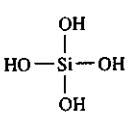
**Registry Number:** 1318-74-7
**CA Index Name:** Kaolinite (Al2(OH)4(Si2O5)) (9CI)
**Other Names:** Kaolinite (7CI, 8CI); Kaolinite (Al2(Si2O7).2H2O); 50R; 50R (mineral); Argirec B 24; Asiacoat HG; ASP 072; Barnett Clay; Biofix C 1; Biofix C 2; Biofix E 1; Biofix E 2; BIOFIX SC; GK; GK (mineral); Hydrite 121; Hydrite flat D; Hydrite MP; Hydrite PX; Hydrite R; Kaopaque 10; Kaopaque 10S; Kaopaque 20; Kaopaque 30; Kaophile 2; Kaophobe 45; Mono 90; SPS Kaolin; UG 90
**Formula:** Al. H O. O5 Si2
**Alternate Formula:** Al2 H4 O9 Si2
**Class Identifier:** Mineral

**Composition:**

| Component | Ratio | Component Registry Number |
|---|---|---|
| O5Si2 | 1 | 20328-07-8 |
| HO | 4 | 14280-30-9 |
| Al | 2 | 7429-90-5 |



**Registry Number:** 14940-68-2
**CA Index Name:** Zircon (Zr(SiO4)) (9CI)
**Other Names:** Zircon (8CI); A-PAX 45M; Standard SF 200; Ultrox 500W; Zircon Flour; Zirconite; Zircosil 15
**Formula:** H4 O4 Si. Zr
**Class Identifier:** Mineral




Zr (IV)





### *SciFinder Scholar Note:*

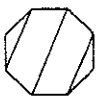
1. Many minerals are easily retrieved through name searches, either directly in REGISTRY or through **Explore by Research Topic**.

### *A5.5.3 Records with 'No References'*

Records with 'no references' may appear because of a number of reasons including:

- manual registration of substances (this may occur because companies need CAS Registry Numbers for commercial purposes);
- CAS registrations for parent ring systems;
- CAS do not use the CAS Registry Number for indexing in the bibliographic database (this particularly occurs for crude natural product extracts); or
- because the substance database may have been updated before the bibliographic database (and hence the CAS Registry Number has been assigned but the complete bibliographic record has not yet appeared).

**Registry Number:** 21426-57-3
**CA Index Name:** Tetracyclo[3.3.0.02,4.06,8]octane (8CI, 9CI)
**Formula:** C8 H10
**Class Identifier:** Ring Parent



**No References**



**Registry Number:** 8001-46-5
**CA Index Name:** Fats and Glyceridic oils, halibut-liver
**Other Names:** Oils, glyceridic, halibut-liver; Oils, halibut-liver; Fish oils, halibut-liver; Halibut oil; Halibut-liver oils; Oils, halibut

(*continued*)

| | |
|---|---|
| **Formula:** | Unspecified |
| **Definition Field:** | Extractives and their physically modified derivatives. It consists primarily of the glycerides of C14-C18 and C16-C22 unsatd. fatty acids. (Hippoglossus hippoglossus). |
| **Class Identifier:** | Manual Registration, Concept |
| | No Structure Diagram Available |
| **No References** | |



### *SciFinder Scholar Note:*

1. 'Substances' of the latter type should be searched under **Explore by Research Topic** through the Index Name.

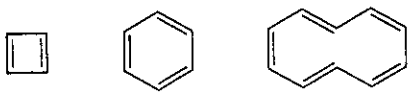
## Appendix 6 Understanding Structure Searches

Structures in REGISTRY have valence bond representations, and hence have single, double, and triple bonds. However, bonds in 'resonance' and 'tautomeric' situations (as defined in the following text) are specified as 'normalized bonds', which effectively means '*either* single or double bonds'.

#### The Resonance Issue

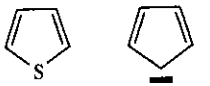
Bonds in substances in which valence bond structures have alternating single and double bonds in rings with *even numbers of atoms* are defined as 'normalized'. Note that the definition does not exactly relate to 'aromatic' compounds since cyclobutadiene (an antiaromatic compound) is defined in REGISTRY with normalized bonds, whereas furan (an aromatic compound) is defined with single and double bonds (since there are an odd number of atoms in the ring).

Examples of structures with normalized bonds:



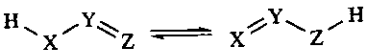


Examples of structures with exact bonds (single and double): -




#### The Tautomerism Issue

Where the following structure representation occurs, the bonds are defined as 'normalized'.




Effectively, X, Y, and Z may be almost any atom, although all three cannot be carbon (simple alkenes are registered with double bonds!). The definition has wide implications and is most commonly encountered with keto/enols and with carboxylic acid derivatives. A key aspect is the hydrogen atom, so, for example, carboxylic acids and primary and secondary amides have 'normalized' bonds, although carboxylic acid esters and tertiary amides have exact single and double bonds.

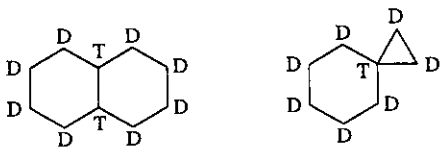
The implications for the searcher are not important at the structure input stage, but often the searcher needs to understand the issue in order to interpret why certain answers have been retrieved. **Analyze References** by **Precision** may need to be used to narrow the answers.

### *Chain-locking Tool*

The structure connection tables in REGISTRY specifically label bonds as either chain or ring bonds. In substructure searches, SciFinder Scholar applies by default a chain *or* ring value to any chain bond drawn in the query. The chain-locking tool overrides the default and ensures that the chain bond drawn only retrieves chain bonds in answers.

### *Ring-locking Tool*

The structure inputs in REGISTRY specifically give all ring atoms an additional label 'D' or 'T'. The 'D' is used when the atom is part of a single ring (i.e. has two ring bonds only), whereas 'T' is used when the atom has three or more ring bonds.



Ring atom descriptions for actual structures in REGISTRY

In structure queries, by default, SciFinder Scholar allows either 'D' or 'T' values to ring atoms drawn with two ring bonds (i.e. left-hand structure *query* would have two atoms 'T' and eight atoms 'D' or 'T'). However, when the ring-locking tool is applied, ring atoms with two ring bonds are specified 'D' values only (i.e. left-hand structure *query* would have eight atoms specified 'D' and two atoms specified 'T').

# Appendix 7 Original Publication Discussed in Chapter 6, Section 6.1

**Bibliography and Abstract**

1676 *J. Am. Chem. Soc.* **1997,** *119,* 1676–1681

## Ester Hydrolysis by a Catalytic Cyclodextrin Dimer Enzyme Mimic with a Metallobipyridyl Linking Group

**Biliang Zhang and Ronald Breslow***

*Contribution from the Department of Chemistry, Columbia University, New York, New York 10027*

*Received October 29, 1996*®

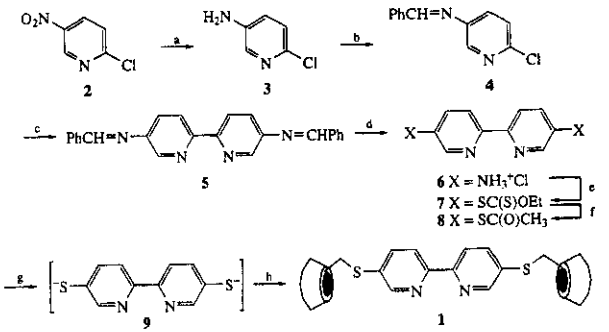
**Abstract:** A β-cyclodextrin dimer with a linking bipyridyl group is synthesized as a catalyst precursor, a holoenzyme mimic. It binds both ends of potential substrates into the two different cyclodextrin cavities, holding the substrate ester carbonyl group directly above a metal ion bound to the bipyridyl unit. The result is very effective ester hydrolysis with good turnover catalysis. For example, a Cu(II) complex accelerates the rate of hydrolysis of several nitrophenyl esters by a factor of $10^4$–$10^5$, with at least 50 turnovers and no sign of product inhibition. In the best case, with an added nucleophile that also binds to the metal ion, a rate acceleration of $1.45 \times 10^7$ over the background reaction rate was observed. Hydrolysis by a catalyst with only one cyclodextrin binding group is significantly slower than in the bidentate binding cases. As expected, the binding of a transition state analogue to these catalysts is stronger with the metal ion present than without. This and kinetic evidence point to a mechanism in which the metal ion plays a bifunctional acid–base role, enforced by the binding geometry that holds the substrate functionality right on top of the catalytic metal ion.



**Schemes 1 and 2**

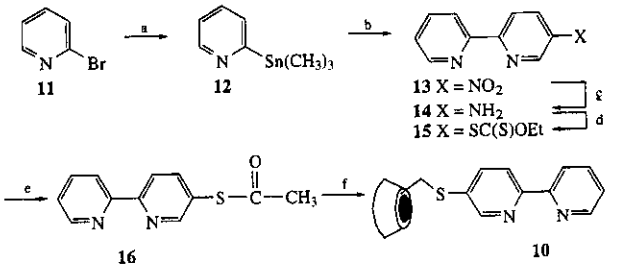
*Catalytic Cyclodextrin Dimer Enzyme Mimic*

**Scheme 1**



(a) Fe/$NH_4Cl$, MeOH, $H_2O$,rt; (b) PhCHO, $MgSO_4$, $Et_3N$, $CH_2Cl_2$, rt, 24 h; (c) $NiBr_2(PPh_3)_2$, Zn/$Et_4N^+T^-$/THF, 50–80 °C, 20 h; (d) 1 N HCl, reflux; (e) (1) $NaNO_2/H_3^+O$, (2) KSC(S)OEt, $H_2O$, 65–70 °C; (f) (1) 20% KOH/EtOH, reflux, (2) $CH_3COCl$, 0–5 °C; (g) $NH_3$/MeOH, rt, 1 h; (h) mono-6-iodo-beta-cyclodextrin, DMF, 60–65 °C, 3h.

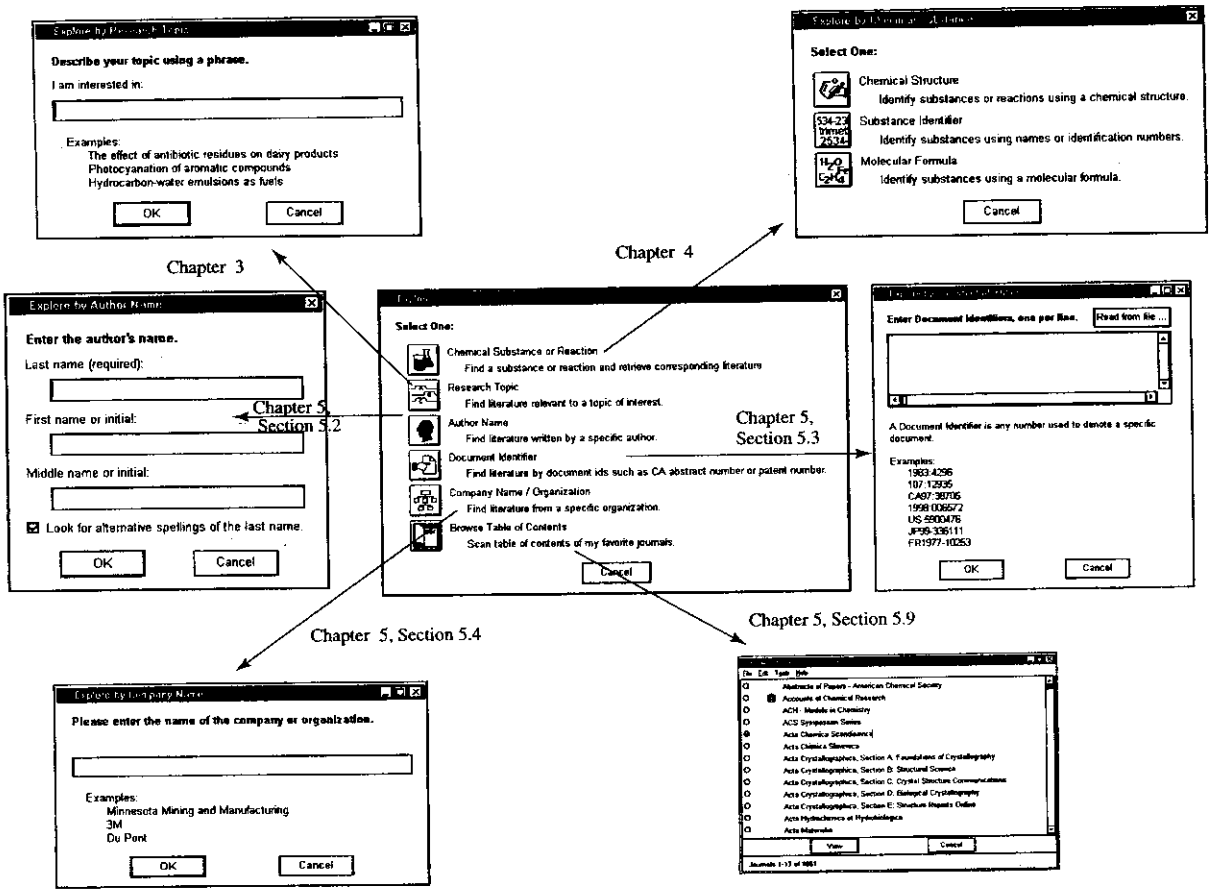
**Scheme 2**



(a) (1) BuLi/$Et_2O$, −78 °C, (2) $(CH_3)_3SnCl$, THF, −78 °C; (b) 2-chloro-5-nitropyridine, $Pd(Ph_3P)_2Cl_2$, THF, reflux, 24 h; (c) Pd° on activated carbon (10%), $NaBH_4$, MeOH, rt, 5 h; (d) (1) $NaNO_2/H_3^+O$, (2) KSC(S)OEt, $H_2O$, 65–70 °C: (e) (1) 20% KOH/EtOH, reflux; (2) $CH_3COCl$, 0–5 °C; (f) (1) $NH_3$/MeOH, rt, 1 h; (2) mono-6-iodo-beta-cyclodextrin, DMF, 60–65 °C, 3 h.

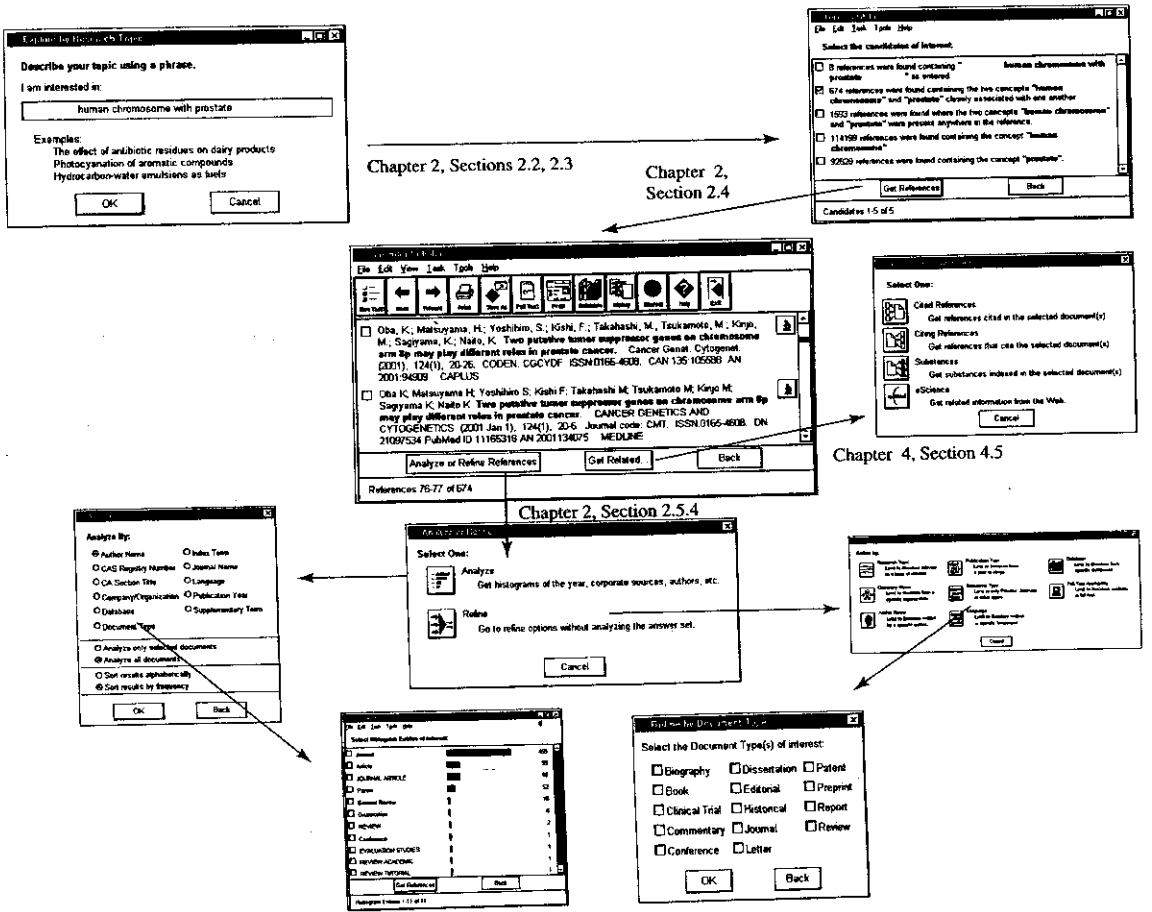
# Appendix 8 Overview of SciFinder Scholar Screens

## A8.1 Working from the Initial Screen

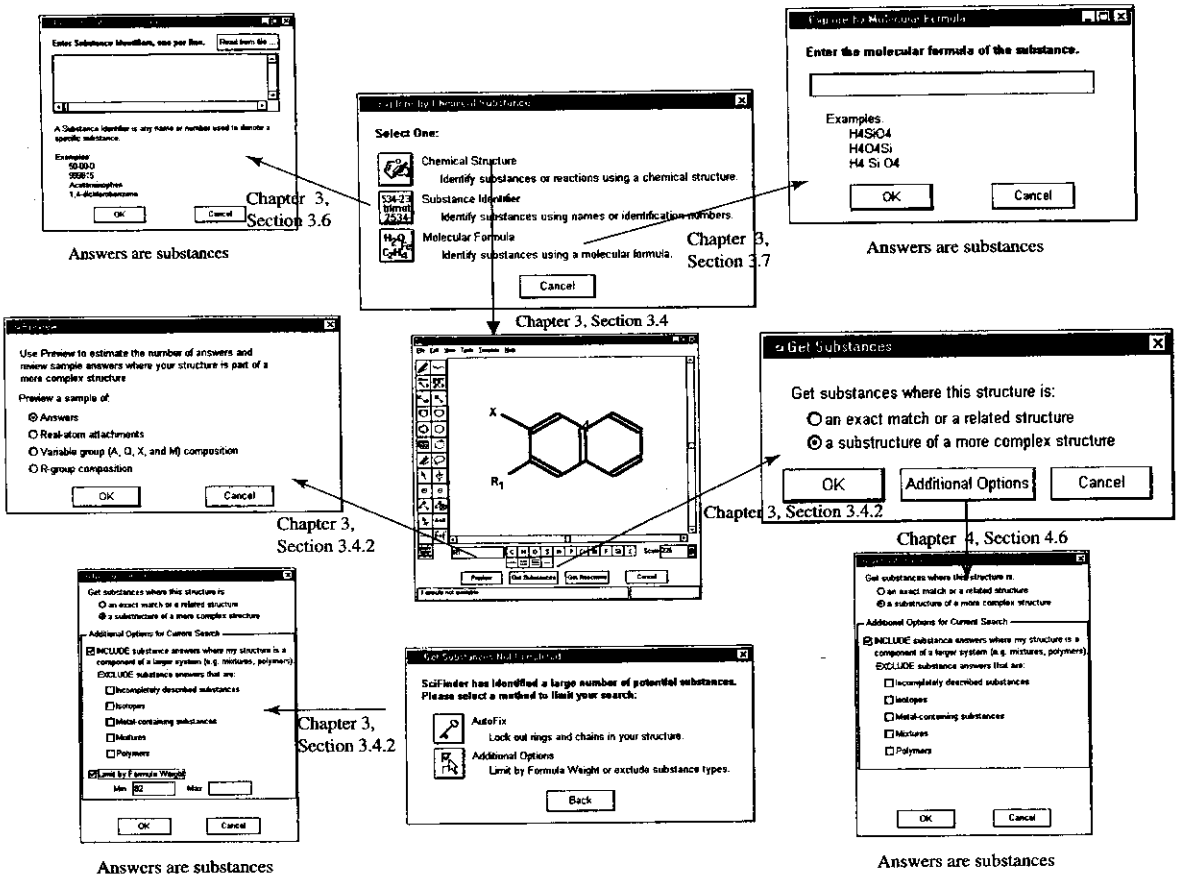





## A8.2 Working from Explore by Research Topic

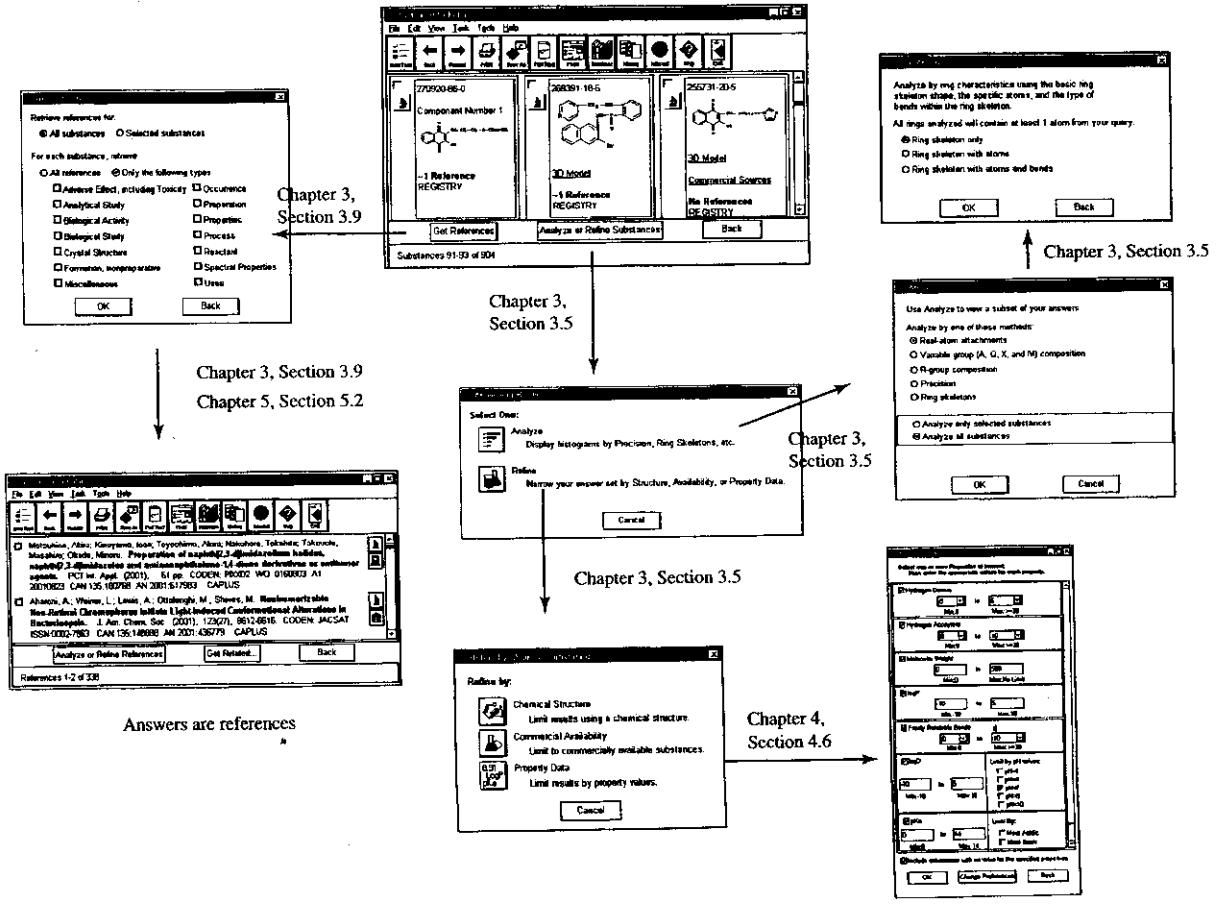


# A8.3 Working from Explore by Chemical Substance

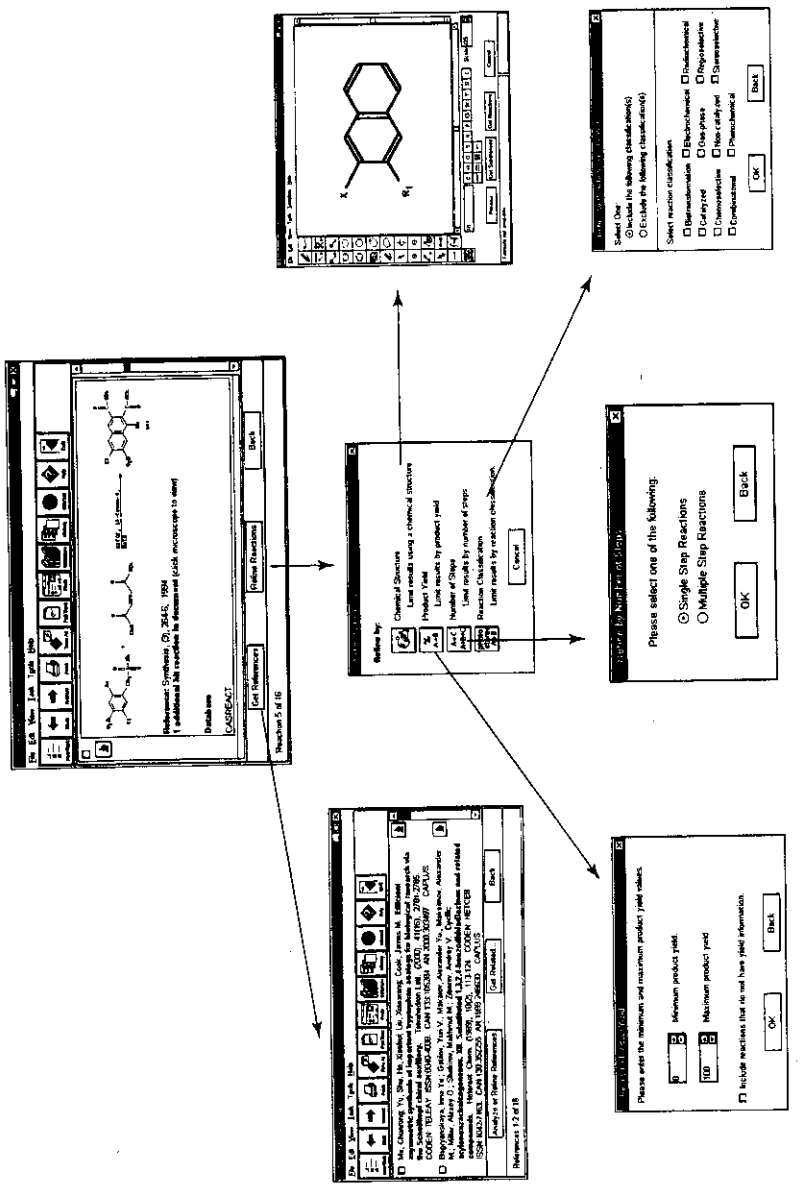





# A8.4 Working from the Substance Answer Screen



**A8.5 Working from the Reaction Screen**




# Index