



# DERWENT WORLD PATENTS INDEX

## **Introduction to Derwent Chemical Indexing**

© 2000 Thomson. All rights reserved

Edition 2  
ISBN: 0 901157 73 2

© 2000 Thomson  
Published by Thomson Scientific  
14 Great Queen Street, London WC2B 5DF,  
United Kingdom

Visit the Thomson Scientific web site at [www.thomsonscientific.com](http://www.thomsonscientific.com)

First edition published September 1997  
Second edition published October 2000

ISBN: 0 901157 25 2 (Edition 1)  
ISBN: 0 901157 73 2 (Edition 2)

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means – electronic, mechanical, recording, photocopying or otherwise – without express written permission from the copyright owner.

Derwent gratefully acknowledges the work of Dr. Paul Bryan, who prepared the bulk of the text for the first edition of the Guide, published in 1997. This second edition has been updated to reflect the changes in Derwent products introduced since the first edition was published.



# Contents

Introduction .....	1
Why is Chemical Code Searching Important? .....	2
Historical Background .....	4
The Logical Basis of the Chemical Codes .....	6
Early Database Technology: Punch Cards .....	12
Searching for a Sample Structure .....	14
Markush TOPFRAG – Code Expert in a Box .....	17
The Chemical Code Search Strategy .....	21
False Drops .....	26
The Bane and Benefit of False Drops .....	28
Customer Support .....	30



# Introduction

This Introduction to Derwent Chemical Indexing answers two basic questions:

- **What are the BCE Chemical Codes?** and
- **Why is chemical code searching important?**

Briefly stated, the BCE chemical codes are a set of alphanumeric symbols, each representing a fragment of a chemical structure, a biological activity, a formulation, a use, or a property described in a patent document.<sup>1</sup> The code G563, for example, represents cyclohexane. "BCE" stands for Derwent's sections B (pharmaceutical), C (agricultural), and E (general chemical). The chemical codes are important to searchers because they comprise the **only** indexing language available for **comprehensively** searching the full spectrum of compounds claimed in most of the world's currently active chemical patents.

The material in this book is an introduction, written to help new and prospective users become acquainted with the history, logical construction, and use of the chemical codes. More detailed information about the chemical codes is

available in the Derwent publication *Chemical Indexing User Guide*.

Reading this introductory book is the first step in a three-step learning process. The second step is a training class on conducting patent searches with the chemical codes, followed by step three, an advanced class on chemical code usage.

<sup>1</sup> For the sake of simplicity, the word "patent" is used throughout this guide to mean patent applications, granted patents and supplementary patent documents.

# Why Is Chemical Code Searching Important?

A variety of databases and searching methods are currently in use for finding chemical patents. A simple chemical substructure search or a search for individual compounds is appropriate in some cases. Most patent searches, however, require that **as many relevant patents as possible** be found. The examples presented in this manual will show why comprehensive chemical patent searches require that a chemical code search be conducted in Derwent's World Patents Index (DWPI) database. The chemical codes can be utilised to find patents that would be **difficult or impossible** to find by any other means, because no patent retrieval system provides such in-depth indexing of the world's chemical patents published before 1987.

In 1987, a new indexing system was developed for the Derwent database, "Derwent World Patents Index Markush", which allowed input and search of both specific and **Markush** chemical structures. This system usually allows more precise patent searches than the BCE chemical codes because it is based more directly on the type of chemical structure frequently found in patents. (The meaning of the term

*Markush*, which refers both to the newer indexing system and the kind of chemical structures found in patents, is explained more fully on page 6). In 1998 Derwent and the French Patent Office, INPI, decided to combine their Markush files into a single service, the Merged Markush Service (MMS) and to index both forwards and backwards in time. Consequently, Derwent is indexing new patents whilst INPI is initially indexing Pharmaceutical patents from WO, EP, FR, DE, GB and US prior to 1987.

In 1999, a new database for searching specific compounds indexed to ongoing DWPI records was released on STN, the DCR structure database. DCR is searchable by chemical structure and also by a number of text fields such as names, molecular formula, molecular weight, etc., and allows quick and simple access to the bibliographic record in DWPI after searching the structures.

However, the newer indexing systems do not supersede the BCE chemical codes. There will be active patents that are not found in the new Markush databases at least until the year 2006, the year that most

1986 European patents will expire. In addition, MMS is only available on the Questel.Orbit online host, whilst the new DCR database is currently only structure searchable on STN. This means that searching with the chemical codes will be necessary for **comprehensive** patent retrieval for many years to come, and learning to use them effectively is well worth the time invested.

# Historical Background

The value and innovation of the chemical coding system can be best appreciated by understanding the historical context in which the codes were first introduced.

The chemical codes were created in the early 1960's out of a need to enhance access to the growing number of chemical patents. Most patent indexes at that time were compiled by choosing the most important compounds described in patent examples, and identifying these **individual compounds** with chemical nomenclature or special compound numbers. General chemical keywords such as *lactam*, *benzotriazole*, and *anthracene* were also assigned to the indexing record, in order to indicate the general class of compounds involved in the patent. A person looking for patent information could then search for individual compounds and general chemical classes in patents using the same nomenclature, compound numbers, and keywords.

As the number and complexity of chemical patents increased, the indexing methods of the day became less and less adequate for **comprehensive** patent retrieval. Patents claiming thousands of compounds became commonplace, and they could not be completely described using chemical terms or identifiers for single compounds.

Compiling such a list of compounds would take too long, and the indexes would become unmanageably large. Another problem was that no indexer could know which compounds claimed in a patent would be important to the world's chemists for its ensuing 17 to 20 year life span.

It was in this environment that Derwent developed a new indexing language that was more suitable for describing chemical patents, called the **BCE Chemical Fragmentation Codes** (also known as "the chemical codes", "the fragmentation codes", or to initiates, simply "the codes"). As its name implies, this indexing language is based on **chemical fragments**, i.e. parts of molecules, present in the description of an invention. Codes for **biological activities, formulations, uses** and **properties** were created as well. Derwent's indexers didn't try to list all of the important chemicals in a patent. Instead they listed all of the chemical fragments that were present in the various molecules covered by each patent. This was a completely different approach from keyword and single compound indexing, and as will be seen later, it greatly increased the flexibility and comprehensiveness of patent searches. Some sample chemical codes and their meanings are illustrated on the following page.

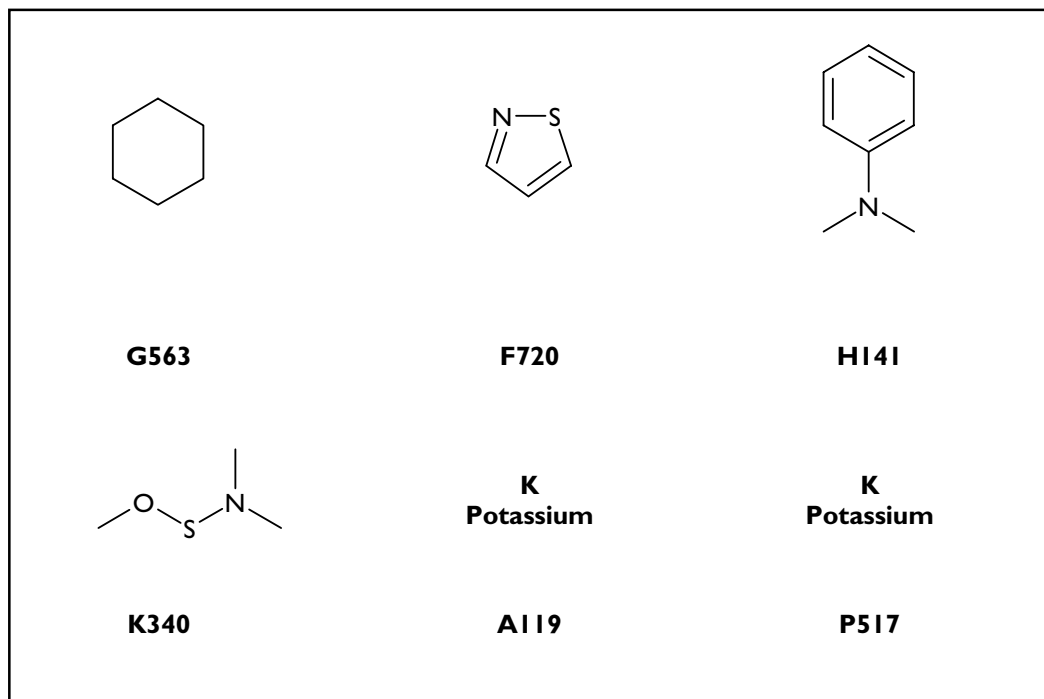


Figure 1 Sample chemical codes and their meanings.

# The Logical Basis of the Chemical Codes

The choice of using chemical fragments for patent indexing was a reasonable one. The fragments were created from the chemical terms used frequently in patents, the structural portions of which often consist of a core ring system with several variable substituents. A few carefully chosen structural fragments can be used to describe almost any molecule and its list of substituents.

Defining a complex variable chemical structure using chemical codes is somewhat analogous to “defining” a dog by giving details about its component parts. Describing the dog’s prominent features, i.e. its ears, nose, hair, body, legs, and tail, indicates the kind of dog it is. Suppose that there were five variations possible for each of the six features; e.g. the ears could be long and thin, short and pointy, short and shaggy, wide and droopy, or long and wide. If each of the six features had five possible variations, it would take a total of  $5+5+5+5+5+5 = 30$  terms to describe all of the possible features; but these **30 possible features** can yield a total of  $5 \times 5 \times 5 \times 5 \times 5 \times 5 = 15,625$  **different combinations!** In the same way, we shall see how listing the chemical fragments present in an invention allows an indexer to describe thousands of possible compounds covered by a patent using a relatively small number of chemical codes.

## Markush structures in patents

Nearly all patents on new chemical entities use what are known as *Markush* structures in the claims. Markush structures are named after Dr. Eugene Markush, whose 1923 patent became a test case for the inclusion of **multiple, independently-varying functional groups** in the description of a chemical invention. In the years preceding Markush’s patent, inventors listed each specific molecule claimed as new; Markush, on the other hand, wanted to patent **categories of compounds**. The first claim of Markush’s patent, US 1,506,316, read as follows:

*“The process for the manufacture of dyes which comprises coupling with a halogen-substituted pyrazolone, a diazotized unsulphonated material selected from the group consisting of aniline, homologues of aniline and halogen substitution products of aniline.”*

Markush won his case, and his victory resulted in a great advance in chemical patent law. Today nearly all new chemical entities in patent claims are described using Markush structures: a **core molecule** that remains constant and **lists of possible**

**substituents.** Figure 2 shows a Markush structure found in US 3,202,699, a 1965 patent assigned to Hoffmann-La Roche.

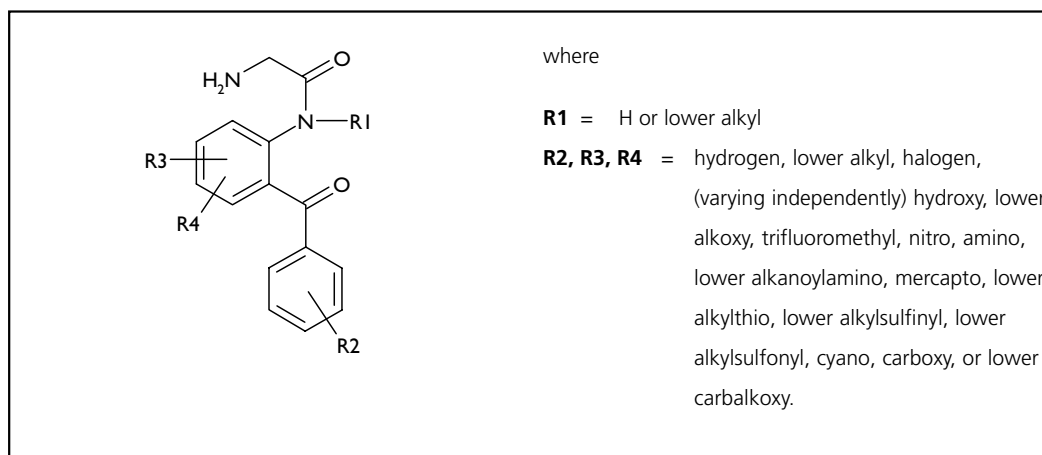


Figure 2 Markush description in US 3,202,699.

## Indexing Markush structures as specific compounds

Doing some arithmetic, it can be determined that the Markush structure in Figure 2 covers more than **150,000 individual compounds**, i.e. it encompasses 150,000 possible variations of the core molecule being patented, even though it mentions **relatively few possible substituents**. Because Markush structures can cover so many individual compounds, it is usually necessary for indexers of “specific compound” databases to choose some of the more important chemicals to include in their indexes. The compounds made in the patent examples are often selected, as well as those that the indexers feel are of industrial significance; keywords are also utilized to describe the broad classes of compounds involved in the patents. Indexers describing US 3,202,699, for example, could have chosen to include structures and keywords similar to those in Figure 3 in their indexing records.

Problems arise with the approach described above when the patent searcher needs to search for compounds that were not indexed, or when the keywords applied are too general for precise, comprehensive retrieval. If one of the many compounds **not** indexed from a particular Markush structure is the subject of a patent search, the corresponding patent is **not**, of course, found in the search results. This can lead to expensive mistakes, such as wasted research money or possibly even an infringement suit by the owner of a previous patent.

Problems with “single compound” patent indexes:

- Patents cover too many compounds to index individually
- Keywords are too broad for precise, comprehensive retrieval

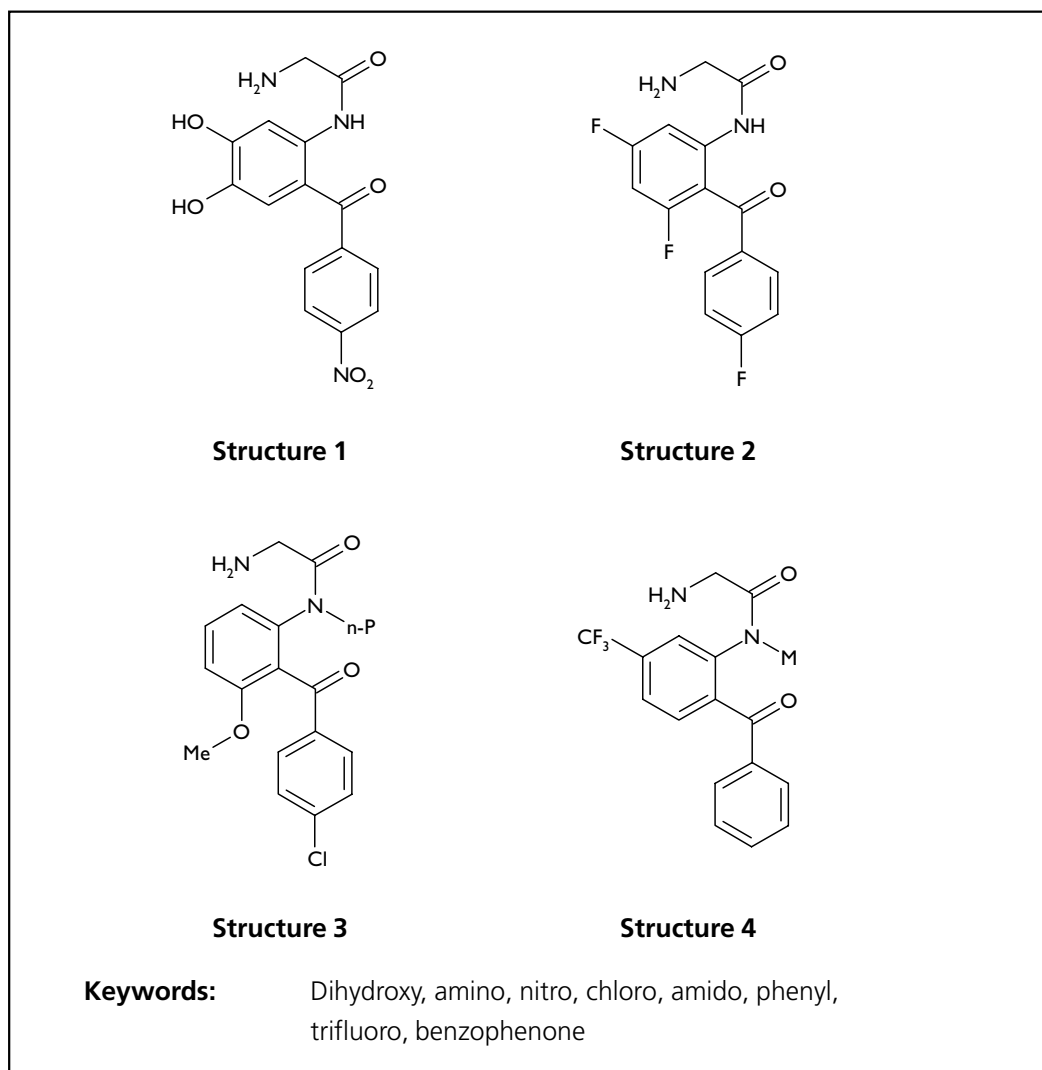


Figure 3 Specific structures and keywords selected from US 3,202,699.

## Indexing Markush structures with chemical fragments

The chemical fragment approach, on the other hand, is quite different. Derwent indexers read each patent carefully to find the widest chemical disclosure, that is, the greatest range of compounds the patent could exclude others from making, using, or selling. **All possible chemical permutations** covered by the patent are separated into appropriate chemical fragments, and these fragments are then translated into chemical codes. As demonstrated earlier, a list of chemical

codes representing all possible rings, elements and substituents mentioned in a patent is much shorter than a complete list of the compounds that could result from all possible combinations of the structural variables. You can verify this statement by trying to sketch all of the molecules defined by the Markush structure on page 6! Shown below are some of the chemical codes that could be used to index that Markush structure; next to the chemical codes are the chemical fragments that they describe. Note that some of the chemical fragments refer to the core structure shown in Figure 2 and some refer to optional substituents.

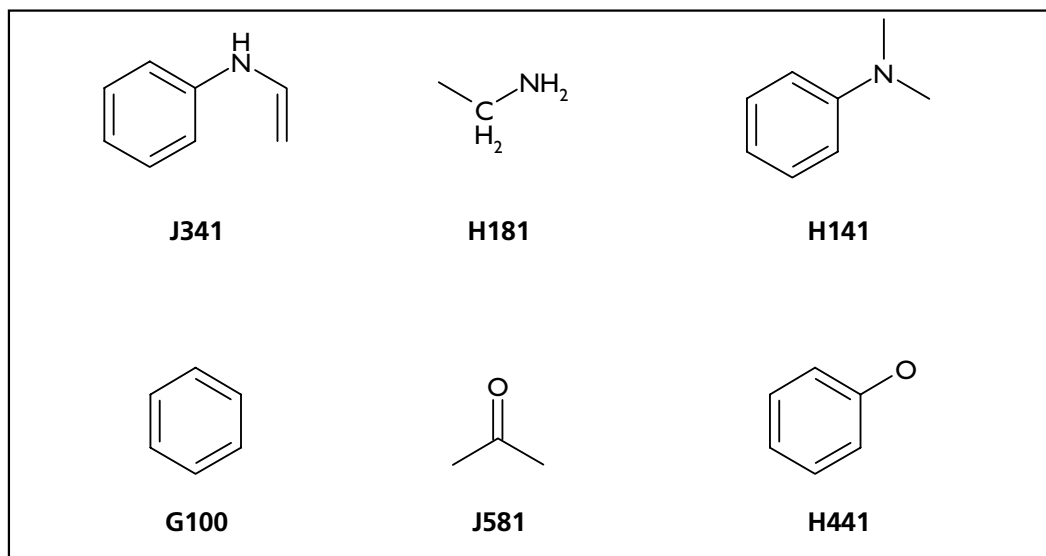


Figure 4 Some of the chemical codes assigned to US 3,202,699.

## Indexing for chemical interest versus indexing the complete coverage of patents

The differences between the various indexing systems that include chemical patents involve more than just the indexing language utilized; the purposes for which the indexes are created may also be quite different. The goal of Derwent's chemical patent indexing system has always been to give users the means by which they can search the **entire range** of compounds claimed by the world's chemical patents. This focus on the complete chemical scope of patents differs significantly from the "new chemistry" focus of other indexing systems that include chemical patents. Indexes that are compiled for chemists usually attempt to point out the major new substances and processes of greatest interest to their users. This is why chemistry-based indexes focus primarily on chemicals actually prepared by the patentee (also known as "wet chemistry"), rather than on the full scope of chemistry claimed in the patent (the so-called "prophetic chemistry"). Derwent's coding, on the other hand, gives no preference to the chemistry in the examples over the chemistry that was claimed but not actually performed, because all of the chemistry

claimed in the granted patent specification is the chemistry owned by the patentee.

Derwent's approach to chemical patent indexing:

- Index the entire range of chemicals claimed by the patent
- Give no preference to chemistry performed versus chemistry claimed

# Early Database Technology: Punch Cards

Computer systems in the early 1960's used Hollerith "punch" cards, i.e. rectangular cards divided into 80 columns and 12 rows, to load data and program statements into the computer's memory. A typical Hollerith card is shown below.

Holes punched at different positions on the card represented different bits of information, and enabled the cards to be sorted by card-sorting machines. Derwent used a modified version of the Hollerith card to represent chemical patent information. Each position on the card,

defined by its row number and column number, was used to represent a chemical fragment, a biological activity, a formulation, a use, or a property found in a patent. For example, if a patent described a new process involving **cyclohexane** or one of its derivatives, cyclohexane was indicated by a hole punched in **column 40, row 2** of the card for that particular patent. A search for all cyclohexane-related patents could be conducted by simply sorting together all cards with a hole punched in column 40, row 2.

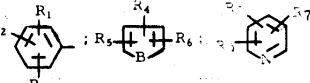
<p>67322R-B. B2. LOVE.13-03-69.          GB-01 402. (+18-07-66/ JB. 36407). R3.          Lovens Kem Fab Produktionsaktieselskab.          C07d (17-08-70).. (pp29).          ALPHA-AMINO-ARYLAMIDETHYL PENICILLIN ESTERS          AS ANTIBIOTICS.. *B1 --7 7140-Q.</p>	<p>B2-P. 1 139          thio, alkylsulphonyl, alkyl, 5-7C cycloalkyl or cycloalkoxy, arylkyl, aralkoxy, aralkylthio; R<sub>1-3</sub> are not all H; R<sub>2</sub>+R<sub>1</sub> or R<sub>2</sub>+R<sub>3</sub>, R<sub>3</sub>+R<sub>4</sub> or R<sub>3</sub>+R<sub>6</sub>, R<sub>6</sub>+R<sub>7</sub> or R<sub>6</sub>+R<sub>9</sub> may form a carbocyclic ring; B is O, S or NH; A is aliphatic, alicyclic, aromatic or heterocyclic radical (opt. subst.) as well as the acid salts thereof and stereoisomers of (I).          Specifically claimed cpds. include: pivaloxy-methyl α-amino-p-chlorobenzylpenicillanate.HCl (II).          (B) Intermediates of formula (IV), below</p>
<p>NEW          (A) Penicillin esters of formula (I)</p> $\text{Ar-CHCONHCH(CH}_2\text{)-CH(S-C(CH}_3\text{)}_2\text{)-CO-N-CHCOOCH}_2\text{OCO(CH}_2\text{)}_n\text{A}$ <p>(where A = </p> <p>n is 0-5; R<sub>1-9</sub> are H, halogen, CF<sub>3</sub>, NO<sub>2</sub>, NH<sub>2</sub>, mono- or dialkylamino, alkanoylamino, OH, alkoxy, alkoxyalkyl-</p>	<p>ADVANTAGES          Antibiotics (I) are absorbed when given by mouth and give very high blood and tissue levels of the corresponding free penicillin.</p> <p>PREPARATION          (1) <math>\text{Ar-CHCOY} + \text{XNHCH(CH}_2\text{)-CO-N-} \rightarrow \text{(I)}</math>          (III)          (where R<sub>10</sub> = NH<sub>2</sub> or gp. convertible thereto by redn. or hydrolysis; COY and XNH are gps. reacting to form CONH).          Contd 67322R</p>

Figure 5 Hollerith card used for data entry in the 1960s

Most chemical patents involve many chemical fragments, so the cards representing those patents had many holes in them. The “**fragments on a card**” approach meant that no matter how many variations of a chemical were claimed in a patent, the Derwent patent card could describe all of them by having holes punched in the appropriate positions on the card. Different chemical inventions within the same patent were usually coded on separate cards. Abstracts were printed on the cards, so cards sorted together could be immediately scanned to judge the relevance of the results. Although punch cards have not been used by Derwent for many years, some people still use terminology like “card records” and “punch codes” when speaking about the chemical codes.

Some subscribers to Derwent’s services bought their own set of chemical code cards. In order to conduct a patent search using the punch cards, metal rods were placed at the positions designated for the chemical fragments and other concepts to be searched, and the cards that matched all of the search criteria were made to drop out of the general stack into an answer stack in the card sorter. That is, the matching answers had all of the right holes. The resulting set of “hits” or “drops” was considered the answer set for the search.

Computer technology gradually became more affordable, and thus more widely available. In 1976, Derwent made the **Derwent World Patents Index (DWPI)** database available to its subscribers through System Development Corporation’s ORBIT search system. Today DWPI consists of over 4 million chemical patent family records that can be searched online through a variety of database vendors.

# Searching for a Sample Structure

Let's take a more detailed look at the chemical codes, this time from the perspective of a search request. Suppose a researcher wants to know if the compound shown below has been previously described in patent documents.

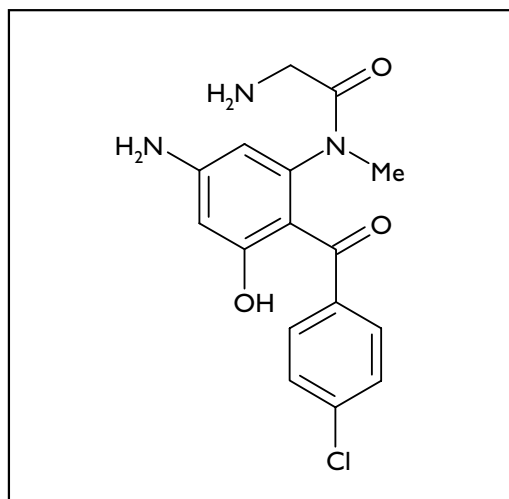


Figure 6 A sample structure search request.

Comparing this structure with the Markush structure of US 3,202,699 on page 7, it can be seen that this compound falls within the scope of the claims of that patent. However, finding patents on this structure in a database is a bit more of a challenge than that.

## Searching with chemical nomenclature

Several methods could be used to search for patents covering the requested compound. Although searches using chemical nomenclature sometimes yield useful results, such searches are often unsuitable for **precise, comprehensive retrieval** because:

- chemical nomenclature varies from patent to patent and database to database;
- most new compound patents mention only a few specific compounds; and
- keywords for broad chemical classes usually find too many results to be helpful in a patent search.

## Searching for specific compounds

Searching specific-compound databases with either compound numbers or a chemical substructure is also a possibility. Searching for chemical substructures is an improvement over the older method of using individual structures or compound numbers because it allows the searcher to search for many chemicals at the same time. However, the indexing records that are

scanned by a substructure search are the same as those that are scanned by a chemical nomenclature search, i.e. a list of individual compounds that were selected by an indexer. The compounds of interest to the searcher may or may not have been included in the indexing records of relevant patents, and thus comprehensive retrieval is not ensured by a substructure search. The cost of missing relevant patents depends on the circumstances prompting the patent search; the search failure could be trivial or it could be extremely costly. If the compound in Figure 6 were a compound the single-compound indexers had selected from the patents that claim it, then it would be a relatively simple matter to find the relevant patents using either chemical nomenclature, a compound number for the requested compound, or a chemical substructure.

### Searching for chemical fragments

Using the BCE chemical codes, a searcher can potentially find all of the patents in which the requested compound was claimed, even if it is buried deep within variable Markush descriptions. This comprehensiveness of retrieval is possible because Derwent indexers attempt to

include **all chemical fragments** from a patent's Markush structure in the indexing record for that patent; thus, practically all possible molecules from each patent are searchable using a list of desired chemical fragments.

To construct a chemical code search on the compound in Figure 6, the searcher first divides the compound into fragments corresponding to BCE chemical codes. Some of the fragments that could be derived from the requested structure are shown below with their corresponding chemical codes.

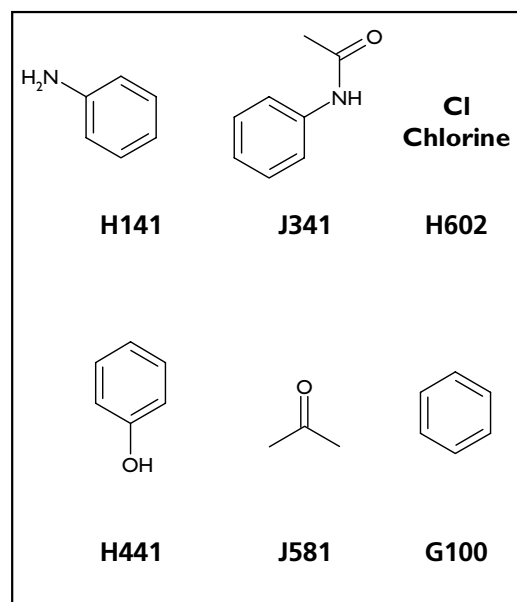


Figure 7 Some of the fragmentation codes used to describe the structure in Figure 6.

In summary, indexes that refer to **individual compounds** allow precise retrieval through chemical keyword and substructure searches, but **only** when the compounds searched are among those that were indexed. Due to practical considerations, the **majority** of compounds covered by Markush structures **cannot** be indexed or listed individually. Listing chemical fragments, on the other hand, allows nearly all compounds specified in Markush structures to be indexed and retrieved.

# Markush TOPFRAG – Code Expert in a Box

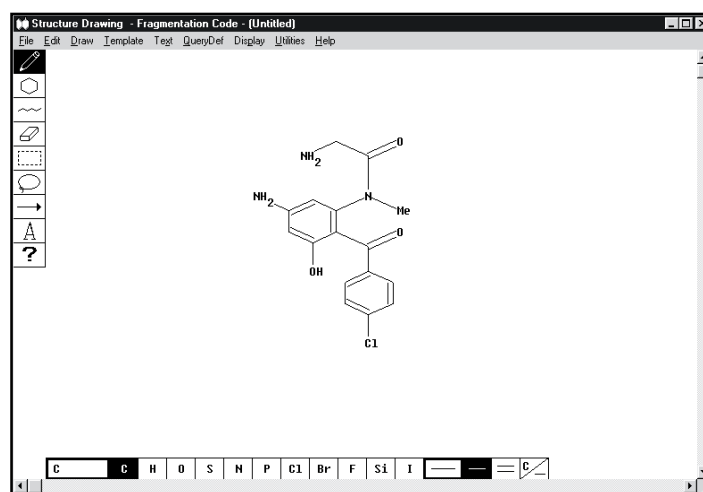
As the memory capacity and processing power of computers increased, it became possible to program more and more intelligence into microcomputer software. In 1987 Derwent introduced TOPFRAG, a microcomputer program that:

- 1 analyzed chemical structures drawn by the user (also known as **topological structures**);
- 2 converted them into the corresponding **fragmentation codes** (hence the name TOPFRAG); and
- 3 constructed an appropriate search strategy that took into account the rules governing the use of the codes.

**Markush TOPFRAG**, the most recent version of the TOPFRAG software, allows the searcher to create substructures that utilize free sites and the general chemical terms that are found in patents, e.g. *alkyl*, *heterocycle*, and *halogen*. With Markush TOPFRAG, a user can learn to use the chemical codes at a convenient pace, and can formulate effective chemical code

searches even during the early stages of the learning process. The search strategies created by Markush TOPFRAG can be simple or very complex, depending on how many variable units are included in the structure to be searched.

We shall use the structure in Figure 6 to briefly illustrate the use of Markush TOPFRAG. First, the user draws the structure to be searched using Markush TOPFRAG's structure-drawing module as shown below.



**Figure 8** Drawing a chemical structure in Markush TOPFRAG.

A few seconds and a few keystrokes later, a search strategy is automatically generated (see Figure 9). The meaning of this search strategy, which looks quite different from most keyword search strategies, is discussed in the section entitled “The Chemical Code Search Strategy” on page 21. Markush

TOPFRAG can be used to create chemical code search strategies for any of the database vendor systems on which DWPI is available, and once formulated, the strategy can be edited and uploaded to DWPI for searching.

```

S M0,M2,M3=(G100(S)H141(S)H181(S)H401(S)H441(S)H602(S)J341(S)J581(S)M150)
S S1(S)M0,M2,M3=(M414(S)M532)
S S2(S)M2,M3=(M121(S)M131)
S S3(S)M2,M3=(M210(S)M281(S)M311(S)M321(S)M342(S)M391(S)(M270+M273))
S S4(S)M2,M3=(M380+M381)
S S5(S)M2,M3=(G013(S)G017(S)H101(S)H641(S)J011(S)M211(S)M349)
S (S2(S)M0=M900)+(S3(S)M2,M3=M901)+(S5(S)M2,M3=M902)+S6
S S7(NOT S)M2,M3=(H2+H3+H5+H7+H9+J1+J2+J4+J9+K0)

Key: (S) = Link operator;
      + = 'OR' operator

```

**Figure 9 Search strategy (Dialog format) generated by Markush TOPFRAG for the structure in Figure 6.**

Among the results of an online search using the chemical codes in Figure 7 would be US 3,202,699, the Markush structure which was seen earlier on page 7. The online

record for US 3,202,699, along with the chemical codes assigned to it, are shown in Figure 10 on the next page. The codes illustrated in Figure 7 are highlighted.

DIALOG(R) File 350:DERWENT WPI  
(c)2000 Derwent Info Ltd. All rts. reserv.

000507534  
WPI Acc No: 1966-08076F/196800  
**Benzodiazepine derivs-sedatives**  
Patent Assignee: HOFFMAN-LA ROCHE (HOFF )  
Number of Countries: 010 Number of Patents: 012  
Patent Family:

Patent No	Kind	Date	Applicat	No	Kind	Date	Main IPC	Week
FR-1329160	A							196800 B
AU-6219646	A							196801
CA-708194	A							196801
CA-725188	A							196801
CH-414655	A							196801
DD-40504	A							196801
DE-1445863	A							196801
GB-985638	A							196801
JP66016380	B							196801
JP68017192	B							196801
NL-280572	B							196801
US-3202699	A							196801

Priority Applications (No Type Date): 61US-0123964 A 19610711  
Abstract (Basic): FR 1329160 A  
Compounds where R1 = H or alkyl. R2, R3, R4 = H or hal. or alk., OH, Oalk, CF3, NO2, NH2, NHCOalk., SH, Salk., SOalk., SO2alk., CN, CO2H, CO2alk.  
Anti-convulsive agents Muscle relaxants Sedatives and tranquillisers IV in AcOH containing 20% HBr is stirred at r.t. for 35 min. Add Et2O rapidly with stirring. Decant., swirl residue with Et2O for 10 min. Decant. Adjust to pH 11 with NH4OH, ext., with CH2Cl2 wash, dry, evap., cryst. hexane-benzene (V).  
Title Terms: BENZODIAZEPINE; DERIVATIVE; SEDATIVE  
Derwent Class: B00  
File Segment: CPI  
Manual Codes (CPI/A-N): B06-D07; B10-A10; B10-A12; B10-B02; B12-C08; B12-C10; B12-D04; B12-E02  
**Chemical Fragment Codes (M0):**  
\*01\* D780 **G100** M533 M532 M531 K442 K499 L460 L499 L140 L199 **H141**  
H181 H142 H143 H211 J131 J132 J133 **J341** J342 H401 **H441** H442  
H443 H444 H494 J521 **J581** J211 J212 H341 H342 H343 H541 H594  
H542 H543 H599 H601 H608 H609 H685 **H602** H600 P446 P448 P447  
P442 P517 M412 M414 M900

Figure 10 The DWPI record for US 3,202,699.

Although new users will find Markush TOPFRAG an invaluable aid, the search strategies it generates may require some editing. Even a beginner should consult the Derwent Chemical Indexing User Guide for the definitions of the codes generated, in order to make modifications that cause the strategy to search more precisely for the information desired. Modifying a search strategy is, in practice, much easier than constructing the search strategy from scratch. Markush TOPFRAG allows the searcher to learn a great deal about the chemical codes and chemical code search strategies offline, when no search costs are involved. Additionally, Derwent makes available a series of classes that teach the meanings of the chemical codes, and how to modify Markush TOPFRAG searches for the best results. For a schedule of current classes, please call the Derwent Help Desk.

In summary, Markush TOPFRAG:

- Allows offline practice in using the chemical codes
- Creates both chemical code and Markush DARC search strategies (see page 29)
- Translates a variable chemical structure into chemical codes
- Creates search strategies that take into account all code rules
- Creates search strategies for any vendor system chosen by the user
- Provides a definition for each code in the search strategy

**Note:** STN Express includes the modules necessary for generating a fragmentation code strategy for use on STN. This is specifically written to facilitate running the strategy using the standard Run Query commands incorporated within the program.

# The Chemical Code Search Strategy

Learning to manually construct chemical code search strategies presents the new user with a number of challenges. Although Markush TOPFRAG handles most of these challenges automatically, this section describes for the interested reader some of the factors that are involved in the chemical code search strategy logic. As you read this section, remember that Derwent's Help Desk staff are available to answer any questions that arise.

## Subheadings

As stated earlier, the number and complexity of chemical patents has increased greatly over the years. To make online patent searching more economical and efficient, Derwent created **subsets** of the patent records found in sections B (pharmaceutical), C (agricultural), and E (general chemical) of DWPI. The subsets correspond to the main BCE chemical subject categories: pharmaceuticals, agriculture, general chemicals, steroids, natural products, dyes, and formulations. The subsets are designated by the **subheadings** M0 through M6 as follows:

M0	Agricultural, Pharmaceutical	1963 - 1969
M1	Agricultural, Pharmaceutical Natural Products and Polymers	1970 to present
M2	Agricultural, Pharmaceutical	1970 to present
M3	General Chemical	1970 to present
M4	Dyes	1970 to present
M5	Steroids	1963 to present
M6	Galenicals	1976 to present

Subheadings can be used in the search strategy to limit the search to the appropriate subset(s) of patents, avoiding the costly process of running a search on unwanted patent categories. The subheadings are actually field labels; when a database record has any of the fields (subheadings) specified in the search strategy, the codes from the search strategy are compared with the codes in that same field in the database record to see if there is a match. Markush TOPFRAG allows the user to specify the subsets of the database to be searched, and then **automatically** creates a search strategy that only runs in those particular subsets. The search strategy generated earlier by Markush TOPFRAG (Figure 9) was created to run in subheadings M0, M2, and M3; according to the subheading list above, these subheadings are used for all pharmaceutical, agricultural,

and general chemical patents that do not involve natural products, polymers, dyes, steroids, or galenicals. The chemical codes in the patent record in figure 10 are preceded by the subheading M0 because this record represents a pharmaceutical patent published in 1965.

### New codes and time-ranging

As the number of patents in Derwent's database increased, the original chemical codes were eventually rendered too general to create small enough sets for a searcher to scan conveniently. Stated in a different way, the number of patents that was found by each code increased greatly, so new codes had to be introduced to enable the creation of more specific search strategies.

**New sets of codes** were added to the existing codes in 1970, 1972, and 1981. In

each instance, the newer codes were only assigned to patents published after they were created. This made it necessary for the searcher to create a different search statement for each effective time period (1963 - 1969, 1970 - 1971, 1972 -1981 Derwent week 26, 1981 Derwent week 27 - present), utilizing the more specific codes only when searching through more recent patents.

Most of the older codes remained in use when new codes were introduced, so search strategies for more recent patents include a mixture of older and newer codes. Table 1 shows a few of the codes made available during each of the time periods, and the time periods during which these codes can be used for searching. For example, in order to search for patents published during the time period 1970 - 1971, the codes introduced in 1963 and the codes introduced in 1970 could be used.

Time period	Sample codes introduced at start of time period	Sample codes searchable during time period
1963-1969	D400, B115, J332	D400, B115, J332
1970-1971	M131, M521, M540	D400, B115, J332, M131, M521, M540
1972-1981/week 26	M313, M343, M131, M521, M540	D400, B115, J332, M313, M343
1981/week 27 to present	D012, H641, L930 M131, M521, M540	D400, B115, J332, M313, M343, D012, H641, L930

Table 1 Sample codes available during each of the four time ranges.

The **standard chemical code search strategy**, illustrated earlier in Figure 9, is compactly constructed to do the work of four different searches, one for each of the four relevant time periods. The search strategy links all of the codes available for searching during each time period (right column in table 1) with a chemical code that represents that time period (left column in table 1).

The chemical codes assigned to the four time periods are:

Chemical Code	Time period
M900	1963 - 1969
M901	1970 - 1971
M902	1972 - 1981 (Week 26)
M903	1981 (Week 27) - present

Table 2 Chemical codes assigned to the four time periods

Every patent record in the DWPI database has only one time-range code in its code list, the one that identifies the time range during which the patent was indexed by Derwent. The patent record in Figure 10, for example, contains the time-range code M900 because it was published in 1965 during the first time range.

Derwent designed a chemical code chart, known as the “coding sheet”, that enables the searcher to quickly determine the year that each of the codes was introduced. A section of the coding sheet is shown below in Figure 11; a complete copy of the coding sheet can be found at the back of this book.

M: MISCELLANEOUS DESCRIPTORS															
M	M1: Rings—C LINKAGES	M1: M1: code essential	M11: Rings linked by bonds	M111: Benzene—benzene	M112: Benzene—aryl	M113: Benzene—other	M114: 'Aryl'—'aryl'	M144: Single heteroatom ex. M141-3	M145: —N—N—	M146: W <sub>n</sub> —In ≥ 2 (ex. M145)	M147: —W—W— (Ws diff.)	M148: —W—W— (ex. M146)	M149: Poly (for M14:)	M150: Aryl-C-aryl	
M26: Chain seq. ex. C≡N	M261# U—S, Se or Te	M262# U=O	M263# U=N	M27: Chain seq. to 9	M271# U—S, Se or Te	M272# U=O	M273# U=N	M340: Valencies all to 1—C att. to ring—C via C=C (excl. M341)	M341: Acyclic—HC	M342: 2—valent excl. M341	M343: 3—valent	M344: ≥4—valent	M349: —C(W)—C(W)— present	M35: Att. to ring —C and V (where V≠X)	M351# Ring—C and V (or CEV) (where V≠X)
M415: Alicyclic	M416: Aliphatic	M417: Incomplete structure	M42: Nat. prod., polymers, apparatus	M421: Antibiotic; vaccine	M422: Vitamin	M423: Other nat. product; polymer	M424: Apparatus	M6: MISCELL. CODES	M610: Compound is hydrocarbon	M620: Saturated aliphatic compound	M630: Metal/amine salt of org. acid	M640: Inorg. acid salt of org. base	M650: Org. acid salt of org. base	M7: PATENT TYPE	M71:—6: Role of compound

Figure 11 A sample section of the coding sheet.

The codes are printed in different colors, which indicates when they were introduced and, consequently, in which time range they belong. For example, the red codes were introduced in 1970. They were the second set of codes to be introduced, and thus are LINKed with the second time-range code, M901. They are also LINKed with the codes from search statement 1, because codes from previous time ranges usually remain valid in subsequent time ranges. Looking at the search strategy in Figure 9

(part shown below), you can see that the codes **M121** and **M131** in search statement 3 are first LINKed with the older codes from search statement 1, and then with the year code **M901** in search statement 7.

```
S M0,M2,M3=(G100(S)H141(S)H181(S)H401
(S)H441(S)H603(S)J341(S)J581(S)M150)
S S1(S)M0,M2,M3=(M414(S)M532)
S S2(S)M2,M3=(M121(S)M131)
. . .
S (S2(S)M0=M900)+(S3(S)M2,M3=M901)
+(S5(S)M2,M3=M902)+S6
```

Finding M121 and M131 on the coding sheet reveals that they are indeed red, which according to the color key on the coding sheet verifies that they were introduced in the second set of codes in 1970. Of course Markush TOPFRAG is able to construct the strategy in Figure 9 correctly without the user having to know anything about time ranges.

## The LINK operator

The chemical code search strategy makes extensive use of the **LINK** operator. The LINK operator creates a “hit” when the terms searched are present together in the same field. In DWPI, chemical codes for

different inventions within the same patent are coded in separate fields; in terms of the older technology, separate inventions within a patent were coded on separate “card records”. If the *AND* operator was used in the search strategy instead of LINK, the codes located anywhere in the same patent record would have generated a hit, even if present in different fields, and hence, in completely different molecules described in the patent. This would have led to the retrieval of many wrong answers.

**Note:** Different hosts use different operators to achieve this 'LINK'ing. Questel. Orbit uses L or LINK, Dialog uses the (S) operator and STN uses the (P) operator.

In addition, if the codes for each time period in the search strategy were not LINKed with the appropriate time-ranging codes, they would have to all be present in a record in order for that record to be retrieved, because all codes are linked together in the search statements before the time-ranging search statement. If the search strategy in Figure 9 had no time-ranging statement, the search would then only retrieve patents published in 1981 week 27 and later, because only those patents could possibly have the codes in search statement 6 (G013, G017, etc.). As shown previously, each time-ranging code is linked with the codes

introduced at the beginning of that time range, as well as with all codes introduced previous to that time range. This is because each patent record with a certain time-ranging code will only contain codes introduced during or prior to that particular time range.

## Code definitions and rules of usage

Learning the precise **definition of each of the codes** enables the searcher to edit Markush TOPFRAG search strategies, fine-tuning them so that they will search more precisely for the type of answers that are desired. Derwent's training classes, user guides and Help Desk staff, described in the section entitled "Customer Support", are available for assisting in this process. Detailed **rules governing the use of the codes**, i.e. the search situations in which each code is used, are found in the Derwent Chemical Code Dictionary. As would be expected, Markush TOPFRAG has been programmed to automatically use the codes in a manner conforming with these rules.

A full explanation of the search logic utilized by the Standard Search Strategy is beyond the scope of this introductory manual; the Derwent Chemical Code

Dictionary discusses it in detail. The purpose of the search strategy, however, is straightforward – to search only the selected subsets of the database, using the most specific codes possible for each of the four time periods that are delineated by the introduction of new codes.

In summary:

- **Subheadings** limit a search to relevant categories of patents
- **Time-ranging** utilizes the most specific codes available in each of the four applicable time periods
- **Code rules and exceptions** define correct code usage
- Markush TOPFRAG can be used to create a **Standard Search Strategy**, which takes all of the above factors into consideration
- Knowing **code definitions** enables the searcher to edit Markush TOPFRAG strategies

# False Drops

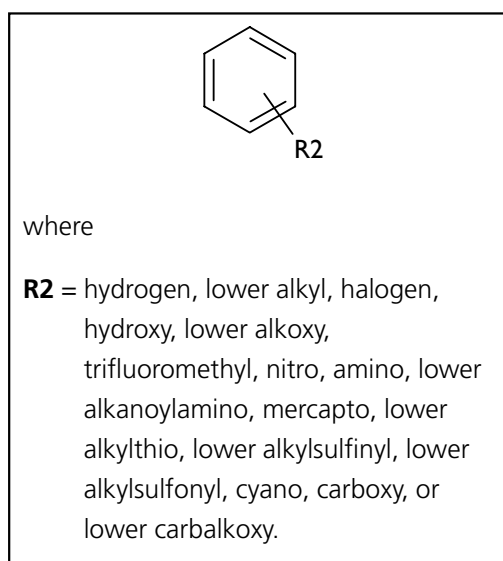
There are usually patent records in the results of a chemical code search that do not claim the exact compound(s) submitted in the search request. Such records are called “false drops”, a term borrowed from the days of card-sorting machines. False drops sometimes occur because the chemical code search strategy allows the searcher to specify the chemical fragments found in the structure to be searched, but not necessarily **how the fragments are connected**. That is, topological information is lost in the translation. Also, some codes represent more than one structural possibility. The code F432, for example, represents both dihydropyridine and tetrahydropyridine. If F432 is used to search for tetrahydropyridine, it may find patents in which F432 was used to represent dihydropyridine.

Another source of false drops comes from fact that all of the fragments from a particular Markush structure are translated into codes, and those codes are listed together in the database record. But the original Markush structure describes **many compounds**, so there are codes listed together that are from the same Markush

structure, but which actually refer to different individual chemical structures. Codes that the searcher intends to be present together in one structure, i.e. the one in the search request, may find results in which those same codes have been derived from different molecules represented by a single Markush structure.

Think for a moment about this process in terms of the old punch cards. First, a Markush structure with all of its possible variations was analyzed using the chemical codes. Then, each relevant code position was punched on the patent's card. No single molecule contained all of the chemical fragments represented by the holes on the card; they were derived from the **complete range of possible molecules** that fell within the scope of the Markush structure. If a search was then conducted for a single molecule having 10 particular chemical fragments, some false drops would probably be in the results, because, although those same 10 codes were punched on every card in the search results, the holes may have been punched for several different molecules described by one Markush structure.

Let's look at a more specific example. The Markush structure from US 3,202,699 on page 7 contains the following fragment:



The Derwent indexer, i.e. the person assigning the chemical codes to the patent, listed the chemical codes for all possible **R2** definitions together in the same indexing record (see the "Chemical Fragment Codes" field in Figure 10). However, the compounds claimed by US 3,202,699 have at most one of these possible substituents on the particular phenyl ring shown above. So if a search is conducted for compounds containing a phenyl ring substituted by a halogen, a hydroxy, *and* an amino group, US 3,202,699 will be one of the hits because it has all the right codes – but there

are no molecules with tri-substituted phenyl rings among the compounds covered by that patent.

In summary, false drops may occur because:

- Chemical codes don't always specify how the fragments fit together;
- Several structural possibilities can be represented by a single code; and
- Codes grouped together in one online record can represent many different molecules.

# The Bane and Benefit of False Drops

The false drops discussed in the previous section, i.e. patent records that don't contain the exact structure for which a search is conducted, can be frustrating. But they can also be very important to a patent case. They may contain compounds similar to, but not exactly the same as, the compound searched. A patent attorney may consider these “**near misses**” to be as important as (or perhaps even more important than) documents that do contain the exact compound searched. The near-miss patents may be a part of the **relevant prior art** of a chemical compound or process that would be **difficult or**

**impossible** to find by other means. The same ambiguity, or “**fuzziness**”, inherent in the chemical codes that allows wrong answers to be part of the search results, also allows **near-misses** to be found, and permits **comprehensive retrieval** of the **complete range of compounds** covered by the Markush structures found in nearly all patents on new chemical entities. For example, reading through the results of a chemical code search for the molecule in Figure 6 one finds GB 1120807 by Smith Kline & French, which contains the Markush structure shown in Figure 12.

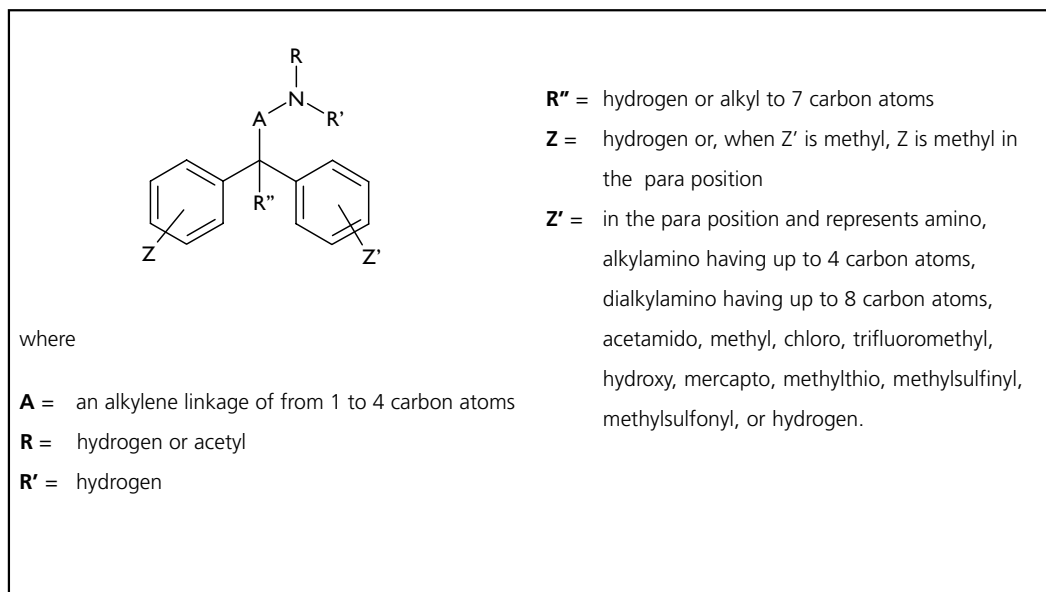


Figure 12 “Near miss” resulting from a chemical code search for the structure in Figure 6.

Although the molecule in Figure 12 has a number of structural differences from the molecule in the search request, a patent attorney may find the technology disclosed in this “false drop” helpful in proving a certain point about the prior art of the compound in question. In other instances, however, the false drops are not useful and are considered “noise”. One of the goals of every aspiring code searcher is to construct searches that retrieve a minimum amount of noise.

False drops may be either:

- **Noise** to be deleted, or
- **Near misses** that form part of the prior art of the compound in question.

## Markush DARC Graphical Indexing

As indicated on page 2, from Derwent Week 8701, both Markush structures and single compounds in Sections B,C and E have been graphically indexed using the Markush DARC indexing language in the MMS file.

Markush indexing allows search results of higher relevance than Chemical Code indexing, because searches are conducted

using chemical structures, that can be either specific or highly variable, unlike the relatively disjointed set of chemical fragments used in Chemical Code searching. Moreover, the Markush DARC search system does not sacrifice comprehensiveness to achieve the increased relevance.

## DCR Graphical Indexing

In the same way, DCR indexing (also described briefly on page 2) allows search results of higher relevance than Chemical Code Indexing. As with other single compound databases, it can also be easier to perform sub-structure searches using DCR (or MMS). However, the retrieval can be lower for the reasons described on page 8.

# Customer Support

## Customer Technical Support

Expert advice and support are available to help answer your individual online enquiries. The Customer Technical Support staff have an in-depth knowledge of the Derwent databases, as well as the command languages of the various online hosts. They will advise you on developing search strategies, incorporating the various specialised capabilities of the host's search systems and explain the results of your search. In addition to subject searching, our Technical Support staff can explain how Derwent classifies technologies in our databases.

### Europe & Rest of World

Thomson Scientific  
14 Great Queen Street  
London WC2B 5DF  
United Kingdom

Tel +44 (0)20 7344 2999  
Fax +44 (0)20 7344 2900  
Email [ts.support.emea@thomson.com](mailto:ts.support.emea@thomson.com)

### North & South America

Thomson Scientific  
1725 Duke Street  
Suite 250  
Alexandria VA 22314  
USA

Tel: +1 703 706 4220  
Free: +1 800 451 3551  
Fax +1 703 838 0450  
Email [custserv@derwentus.com](mailto:custserv@derwentus.com)

### Japan

Derwent Information Japan  
Thomson Corporation Japan Ltd  
5F East, Palaceside Building  
1-1 Hitotsubashi 1-Chome  
Chiyoda-ku  
Tokyo 100-0003  
Japan

Tel +81 3 5218 6500  
Fax +81 3 5218 7840

### Derwent Web Site

<http://www.derwent.com/>  
<http://www.derwent.co.jp/>

## Online Training

Specialist staff run regular training programmes to help you to search our online databases effectively and comprehensively. Training classes are held in major towns and cities in Europe, North America and Japan.

A wide range of classes are available introducing Derwent files to both beginners and experienced searchers and subject specialist classes. These classes include specialist training covering indexing available on Derwent's databases. We can also develop training classes or presentations tailored to your company's specific needs.

## Derwent International Patent Copy Service – ordering patent documents

Having completed your online search you can order quality copies of patent documents issued around the globe. As holders of the world's largest private collection of international patents, Derwent provides a fast and efficient service. In addition, through a global network of contacts, Derwent regularly locates and

supplies old and unusual patents. To make use of this service, simply contact your local Derwent office HelpDesk (see page 30 for details).

## Derwent Search Service – help at hand when you need a search, but lack the time or expertise in-house

When you need an information search, but lack the time, resources or expertise to conduct it inhouse, you can rely on the Derwent Search Service to find the details you need. Using our patent expertise and global network of patent sources, Derwent's Search team will ensure that you receive accurate, timely and confidential search results – in a time-scale and format defined by you.

Our patent and scientific experts can also prepare a detailed value-added analysis of your search results to assess patentability, validity and infringements of other patents.

For more information and details of the service costs simply contact your local Derwent office HelpDesk (see page 30 for details).

## User Guides

- **The Introduction to Derwent Chemical Indexing** is an introductory book that helps new and prospective users of the chemical codes to become acquainted with their history, logical construction, and use.
- **Chemical Indexing User Guide** contains the most detailed information available on the meaning and use of the chemical codes. It is arranged in the order of the codes, from A to V and is an alphabetical subject listing that refers the user from scientific concepts and nomenclature to the corresponding chemical codes.
- **Chemical Code Dictionary** is an alphabetical subject listing that refers the user from concepts and nomenclature to the corresponding chemical codes.
- **Markush DARC User Guide** discusses Markush structure concepts and teaches users how to access and search Derwent World Patents Index Markush database using Markush DARC search software.

- **Chemical Indexing Guidelines** gives advanced users an insight into how Derwent indexes patents.
- **Chemical Coding Sheet** is a concise summary of the BCE Chemical Codes. It gives an abbreviated description of each code, provides an overview of the logical sequence of the codes, and indicates the year that each code became available for searching.

## The Chemical Coding Sheet

A copy of the chemical coding sheet can be found at the back of this book. It is designed to:

- give an abbreviated description of each code;
- provide an overview of the logical sequence of the codes; and
- indicate the year that each code became available for searching.

## Markush TOPFRAG

Markush TOPFRAG is a microcomputer software program that creates chemical code and Markush DARC search strategies from drawn chemical structures. It runs on all standard Windows based operating systems.

Markush TOPFRAG is discussed in the section entitled *Markush TOPFRAG - Code Expert in a Box*.

