Strategies for Database Dereplication of Natural Products

David G. Corley, and Richard C. Durley

J. Nat. Prod., **1994**, 57 (11), 1484-1490• DOI: 10.1021/np50113a002 • Publication Date (Web): 01 July 2004

Journal of Natural Products
Vol. 57, No. 11, pp. 1484-1490, November 1994

STRATEGIES FOR DATABASE DEREPLICATION OF NATURAL PRODUCTS

DAVID G. CORLEY* and RICHARD C. DURLEY

Monsanto Company, 700 Chesterfield Parkway North, St. Louis, Missouri 63198

ABSTRACT.—The rapid characterization of known compounds (dereplication) has become an important consideration for the natural products chemist screening for pharmaceutical or agricultural compounds. Several commercial databases have been developed to assist the chemist in identifying known compounds with minimal amounts of physical or biological data. Strategies have been developed to efficiently search STN (Scientific and Technical Network) files with formula weight, carbon count, structure fragments, bioactivity, and taxonomy. A brief review is also provided on other commercially available databases that are useful for natural products dereplication.

Screening natural product extracts for novel bioactive compounds is an increasingly challenging area of research for pharmaceutical and agricultural companies. The chances of finding novel bioactive compounds have become more difficult due to the enormous number of known compounds already described in the literature. The rapid characterization of known compounds, a process known as dereplication, has become a strategically important area for the natural products chemist involved in screening programs. Several commercial databases have been developed that can assist the natural products chemist in reducing structure elucidation time on known compounds. These databases can be searched with minimal amounts of physical and/or biological data. Chemical Abstracts Service's Registry File, available on the Scientific and Technical Network, STN International®, is the largest online repository of natural product structures. To overcome the sheer size of the Registry File (in excess of 12 million compounds) strategies have been developed to efficiently search using formula weight, carbon count, structure fragments, bioactivity, and taxonomy. Other commercially available databases for natural products dereplication are: Chapman & Hall's *Dictionary of Natural Products, Bioactive Natural Products Database* (also known as The Bérdy antibiotic database), DEREP, MARINLIT, and the *Marine Natural Products Database* (MNP Database). The searchable attributes of these databases are listed in Table 1.

STN International is a collection of over 180 different databases (files) marketed by Chemical Abstracts Service (CAS) in Columbus, Ohio; FIZ ,Karlsruhe, Federal Republic of Germany; and the Japan Information Center of Science and Technology (JICST), Tokyo, Japan. The STN files of particular interest to the natural products chemist are the CA[1], REGISTRY, NAPRALERT, BEILSTEIN, SPECINFO, MEDLINE, EMBASE, JICST, and BIOSIS. Of universal importance to linking information between STN files is the CAS Registry Number®. Once a desirable query is obtained, it can be used in many STN files without further modification, a process known as file crossover. This provides a powerful tool for linking together structure, physical properties, and biological information. Successful searching requires the user to have a good working knowledge of the Messenger℠ language for creating queries. Several good sources and workshops are available to learn the necessary techniques (1–6). In addition, front-end software such as STN Express® greatly reduces the burden of syntax problems, especially for structure searching, and brings the often cryptic world of online searching into the hands of the bench chemist.

[1]The HCA file can be substituted for the CA file to limit cost when a large number of search terms is used.

TABLE 1. Searchable Attributes of Commercially Available Databases for Natural Products Dereplication.

Database	No. of Compounds[a]	UV	FW	MF	Bioactivity	Taxonomy	SSS[b]
STN Files[c]	12,000,000	+[i]	+	+	+	+	+
Ch&H[d]	90,000	−[j]	+	+	+	+	+
BNPD[e]	23,000	+	+	+	+	+	−
DEREP[f]	7,000	+	+	+	+	+	−
MARINLIT[g]	6,000	+	+	+	+	+	+
MNP[h]	4,000	−	+	+	−	+	−

[a]Approximate figures.

[b]SSS=substructure searching.

[c]2540 Olentangy River Road, P.O. Box 3012, Columbus, OH 43210-0012 [tel. 800-848-6533].

[d]2–6 Boundary Row, London SE1 8 HN, UK [tel. 071-865-0066].

[e]H-1808 Budapest, Hungary [tel. (361) 14-22-796].

[f]3150 Rumsey Drive, Ann Arbor, MI 48105-1466 [tel. (313) 665-7171].

[g]D. John W. Blunt and Murray H.G. Munro, Department of Chemistry, University of Canterbury, Private Bag 4800, Christchurch, New Zealand [FAX (643) 364-2110].

[h]John Faulkner, University of California, San Diego, La Jolla, CA 92093-0212 [tel. (619) 534-4259].

[i]BEILSTEIN and HODOC files only.

[j]Chapman & Hall are currently indexing uv data for a future release.

STRATEGIES AND EXAMPLES

The key to successful searching is an understanding of how articles are indexed. Information that is important to a natural products chemist may not be judged novel for indexing purposes at CAS. The lambda-max from a uv spectrum,[2] for example, is not indexed by CAS in the registry file, but is one of the first pieces of information the chemist has available for dereplication—typically from analytical hplc coupled with diode-array detection. When an article describing a new natural product is reviewed by CAS, key words (controlled vocabulary) from the title, abstract, and text are indexed. In addition, producing-organism taxonomy as well as compound attributes are indexed including the trivial name, CAS name, molecular formula, and formula weight (Tables 1 and 2).

TABLE 2. CAS Section Codes for Indexing Natural Product-Producing Organisms and Structure Classes for the CA File.

Divisions of Taxonomy		Structure Classes	
Section Code[a]	Description	Section Code	Description
10	Microbial[b]	26	Biomolecules[c]
11	Plants	30	Terpenes and Terpenoids
12	Nonmammalian	31	Alkaloids
		33	Carbohydrates[d]
		34	Amino Acids, Peptides, and Proteins

[a]Section Codes have been changed over time and may result in missed answers.

[b]From the 13th Collective Index, "Microbial," which includes algal and fungal biochemistry (previously macroalgae and macrofungi in sec 11).

[c]Includes macrolides, flavonoids and β-lactams.

[d]Includes nucleosides and nucleotides.

[2]Uv data are searchable on BEILSTEIN (ca. 15,700 records for natural products) and to a limited extent on the HODOC file.

As a general philosophy, it is important to generate a comprehensive strategy when dereplicating on STN. If an insufficient answer set is obtained from a single search, the same question should be asked in different ways. For example, one strategy is to create a subset of unique natural products (Example 1) and use this subset for further searches (Example 2). This approach is important for limiting the answer set to a reasonable number when asking questions that could contain thousands of answers, e.g., formula weight. However, one should be apprised that limiting an answer set may result in the loss of valid answers.

EXAMPLE 1. Creating a Natural Products Subset.

```
⇒File HCA
⇒Search (26 or 30 or 31 or 32 or 33 or 34)/SC,SX and (Molecular Structure/CV or Structure/IT)
      L1 44,449
⇒Search Natural Product#/IA and (Molecular Structure/CV or Structure/IT or New or Novel or
Isolated)
      L2 33,671
⇒Search (Antibiotic#/IA or Alkaloid#/IA or ?Terpen?/IA) and (Molecular Structure/CV or Struc-
ture/IT or Natural Product# or New or Novel or Isolated)
      L3 36,518
⇒Search L1 or L2 or L3
      L4 76,703
⇒Save L4
```

L1 contains the CA section codes (SC) and cross-referenced section codes (SX) for natural products and links them with the controlled vocabulary (CV) term "molecular structure" and index term (IT) "structure." This strategy will identify references where the structure or structural features of a natural product was determined, but does not eliminate all articles that deal with the synthesis, biosynthesis, analysis and manufacture of natural products. L2 searches both the basic index and the abstract index for the phrase "natural product(s)" and links this phrase with articles describing structure elucidation. L3 is similar to L2 except that the specific natural product class "antibiotic(s) or alkaloid(s) or terpenes" must be contained in the basic index or abstract index.[3] L4 uses the Boolean operator "or" to eliminate duplicated answers that are contained in L1 to L3. By using the Boolean operator "not" [e.g., search L1 not (L2 or L3)], one can determine the unique contribution of each query to the answer set L4 (27, 16, and 20%, respectively) and it becomes clear that a comprehensive natural products subset cannot be obtained by one query. By saving L4, the three line query can be activated for use without retyping. Due to system limitations, it is not known how many unique natural products are described in L4, but perusing a cross section of the answer set it is estimated that 75% of the answers are relevant for dereplication with the others being primarily synthesis or analysis papers. Since some articles describe several new natural products, L4 may very well approach the ca. 65,000 compounds that Chapman & Hall's *Dictionary of Natural Products* is estimated to contain.[4]

Example 2 shows how to query formula weight,[5] carbon count, and taxonomy

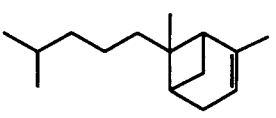
[3]The compound classes "antibiotic, alkaloid, and terpene" are important due to historical reasons. Other compound classes may be used when added emphasis is required. Note that "?terpen?" allows for both left and right truncation which will include the "mono-," "sesqui-," etc., prefixes and the "-oid" suffiix.

[4]Chapman & Hall's CD-ROM database contains more than 90,000 compounds but only an estimated 65,000 are natural products. The remaining compounds are derivatives and related compounds.

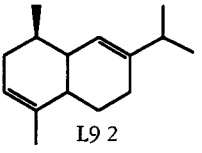
[5]If uncertainties exist in the formula weight, a range search may be more appropriate. Also, it is important to note that only the free-base formula weights are used for alkaloids.

EXAMPLE 2. Searching Formula Weight, Carbon Count and Taxonomy Against a Natural Product Subset.

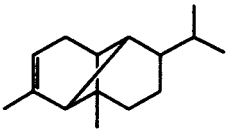
⇒File Registry
⇒Search 204/FW
L5 44,092
⇒Search 15/C and L5
L6 2,161
⇒File HCA
⇒Search L4 and L6
L7 1,120
⇒Search L7 and (zygophyllaceae or bulnesia)
L8 1
⇒Select Hit L8 RN
E1–E5 Assigned
=File Registry
⇒Search E1–E5
L9 5
⇒Display L9 1–5



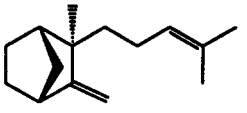
L9 1



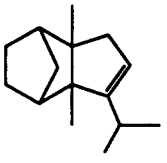
L9 2



L9 3



L9 4



L9 5

against the subset of natural products. As seen in L5, a large answer set of 44,092 compounds is obtained by searching just the formula weight alone. This answer set is too large to crossover in one step and would have to be parsed. L5 was reduced to a more manageable size by using the element count. It was suspected that the compound of interest was sesquiterpene (15/C) and this limited the answer set to 2,161 compounds. If one had structure fragments available, L5 could be used to substructure search against in the Registry File, *vide infra*. By combining L6 with L4, 1,120 articles are describing C_{15}-containing natural products with the formula weight of 204. When using taxonomy as a query it is important not to be too restrictive. In the case of L8, both the plant family and genera were used and not the plant species. Example 2 is used to illustrate the tools available for dereplication, but optimal strategies will vary greatly with each problem.

As of March 1993, the Natural Products Alert Database, NAPRALERT, a database dedicated solely to natural products has been added to STN. This outstanding source of natural product information contains over 104,000 compound records covering ca.

70,000 unique natural products.[6] NAPRALERT indexing is built around numeric and textual classification codes which describe pharmacological activities. NAPRALERT also indexes taxonomy in a very useful format including: class, family, genus, species, organism part, and geographical area. Physical properties of compounds such as molecular formula, molecular weight, or structure fragment cannot be searched directly in the NAPRALERT file and must be done in the Registry File, which can then be easily cross-referenced to NAPRALERT via the CAS Registry Number. A second approach to building a natural product subset is to use the Locator Code (LC) for NAPRALERT in the Registry File (Example 3). The subset L1 can now be used to search against with molecular formula, molecular weight, or structure fragments. As a cautionary note, it should be mentioned that the L1 answer set shows that only 35,157 of the compounds indexed in NAPRALERT have CAS Registry Numbers.

EXAMPLE 3. Using NAPRALERT Locator Code for Natural Products Subset.

```
⇒File Registry
⇒Search NAPRALERT/LC
      L1=35,157
⇒Search 204/FW and L1
      L2=230
```

It is beyond the scope of this paper to discuss substructure searching in detail, but it is one of the most powerful aspects of the Registry and BEILSTEIN files.[7] Example 4 illustrates how very simple fragments obtained from COSY nmr data can be searched to dereplicate the plant phenolic-terpene, bakuchiol. Searching four simple fragments in a single query results in a surprisingly low answer set. The thirteen answers in L2 could easily be inspected for a match with the proton nmr, but was reduced to a single answer by employing the "napralert/lc strategy" used in Example 3.

The BEILSTEIN file indexes natural products that have been isolated in preparative quantities and is searchable with the INP (Isolation from Natural Product) data field. This field includes natural products from microbial, plant and animal sources, as well as coal and oil. Also indexed in the INP field are the syntheses of natural products. Unlike the CA file, BEILSTEIN does not have a post-1966 record limitation, and thus provides a historically continuous database. The BEILSTEIN file also allows numeric searching of up to six uv λ max values, although only ca. 15,700 records contain uv data. Example 5 illustrates how a subset of 61,472 natural product records is created from a file of over 5.7 million compounds. This subset can then be searched by a formula weight range of 465 to 466 mass units to a subset L2 containing 33 records. Finally, a uv λ max range of 285 to 295 nm reduces L2 to the Olivoretin class of chemistry.

DISCUSSION

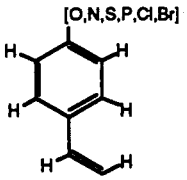
The successful use of STN databases for dereplication of natural products can result in considerable savings in time and money. It is estimated in our laboratory that for each natural product dereplicated, at an average cost of $300 of online time, a savings of $50,000 is incurred in isolation and identification time. Early dereplication also has the added benefit of focusing more resources for the discovery of novel bioactive compounds.

[6]Estimated number of unique natural products by NAPRALERT staff.

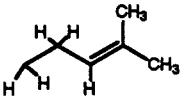
[7]Since substructure searching is expensive it is important that one fully understands the significance of bond and atom attributes of their query.

EXAMPLE 4. Substructure Searching.

⇒Upload Structure
L1 Structure Uploaded
⇒Display L1



Fragment 1



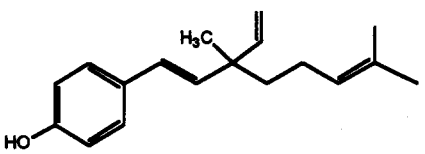
Fragment 2

Fragment 3

CH_3

Fragment 4

⇒Search L1 SSS Full
L2 13 SEA SSS Full L1
⇒Search L2 and NAPRALERT/LC
L3 1
⇒Display L3
L3 ANSWER 1 OF 1 COPYRIGHT 1993 ACS
RN 10309-37-2 REGISTRY
CN Phenol, 4-(3-ethenyl-3,7-dimethyl-1,6-octadienyl)-, [S-(E)]-(9CI)
(CA INDEX NAME)
OTHER CA INDEX NAMES:
CN Bakuchiol (7CI)
OTHER NAMES:
CN (+)-Bakuchiol
CN Drupanol
DR 1408-17-9
MF C18 H24 O
LC BEILSTEIN, BIOBUSINESS, BIOSIS, CA, CAOLD, CAPREVIEWS, CASREACT, EMBASE, MEDLINE, NAPRALERT, RTECS
DES 1:S2:



Cost advantages for online searching of STN databases versus purchasing other commercial databases depends on how often they are used and the need for databases which index all sources of natural products. The searchable attributes of other commercially available databases which are useful for natural products dereplication are listed in Table 1. Chapman & Hall's *Dictionary of Natural Products* is available on CD-ROM and claims to be 95% comprehensive, covering over 90,000 natural products and related compounds. The CD-ROM operates in the Microsoft Windows® environment and features substructure searching. Unfortunately, the lack of searchable uv data prevents

EXAMPLE 5. Searching BEILSTEIN File Using Formula Weight and UV Data.

```
⇒File Beilstein
⇒Search INP/FA
L1    61472 INP/FA
⇒Search L1 and 465–466/FW
      4353 465–466/FW
L2    33 L1 AND 465–466/FW
⇒Search L2 and 285–295/EAM and 230–240/EAM
L3    2 L2 AND 285 NM–295 NM/EAM AND 230 NM–240 NM/EAM
⇒Display L3 1–2 CN
L3 ANSWER 1 OF 2 COPYRIGHT 1994 Beilstein
Synonym (SY):    Olivoretin E
L3 ANSWER 2 OF 2 COPYRIGHT 1994 Beilstein
Synonym (SY):    Olivoretin A
```

this database from being preeminent for dereplication purposes. The Bioactive Natural Products Database [the Bérdy database of antibiotics (7)] is now available for MS-DOS computers. The strength of this database is its indexing of uv (including acid/base shifts) as well as all other important dereplication fields with the exception of substructure searching. DEREP contains ca. 7,000 compounds from mixed phylogeny with 76% of its records citing publications since 1985. DEREP runs on a MS-DOS computer and indexes all the essential fields except for substructure searching. DEREP also does a good job of cross-referencing to analogues and derivatives. MARINLIT is a specialty database focusing on marine natural products but does contain some terrestrial microbes. It runs on a Macintosh computer and is very comprehensive in its indexing. Structures for the ca. 6,000 compounds are available for online display and are substructure searchable with CSC ChemFinder® software. The Marine Natural Products Database (MNP) is Prof. D. John Faulkner's (University of California, San Diego) personal database which is used to write reviews in the journal *Natural Products Reports*. MPD is a Macintosh-based database containing ca. 4,000 records and can be queried for molecular weight, molecular formula, and taxonomy.

ACKNOWLEDGMENTS

We thank Dr. Chris Beecher of the University of Illinois at Chicago for helpful discussions concerning the NAPRALERT database. We also thank Ms. Carol Cochrane and Mr. Dave Deacon of *Chemical Abstracts* for help in demystifying indexing procedures for natural products.

LITERATURE CITED

1. H. Schulz, "From CA to CAS ONLINE," VCH Verlagsgesellschaft mbH, Weinheim, 1988.
2. A. Barth, "The Beilstein Online Database," ACS Symposium Series 436, Ed. by S.R. Heller, American Chemical Society, Washington, D.C., 1990, Chapter 3.
3. "CA File for Chemists," Chemical Abstracts Service, Columbus, Ohio, 1985.
4. "The CA File," Chemical Abstracts Service, Columbus, Ohio, 1985.
5. "CA File Basics," Chemical Abstracts Service, Columbus, Ohio, 1989.
6. "The Registry File Database Description," Chemical Abstracts Service, Columbus, Ohio, 1988.
7. M. Bostian, K. McNitt, A. Aszalos, and J. Bérdy, *J. Antibiot.*, **30**, 633, 1977.

Received 8 December 1993