The Role of Databases in Marine Natural Products Research

6

John Blunt, Murray Munro, and Meg Upjohn

Contents

J. Blunt (✉) • M. Munro
Department of Chemistry, University of Canterbury, Christchurch, New Zealand
e-mail: john.blunt@canterbury.ac.nz, murray.munro@canterbury.ac.nz

M. Upjohn
Physical Sciences Library, University of Canterbury, Christchurch, New Zealand
e-mail: m.upjohn@gmail.com

E. Fattorusso, W. H. Gerwick, O. Taglialatela-Scafati (eds.),
Handbook of Marine Natural Products, DOI 10.1007/978-90-481-3834-0_6,

Abstract

Access to suitable databases is essential for the rapid dereplication (assessment of novelty) of crude extracts in natural product research. The efficiency of this dereplication process relies heavily on the interpretation of molecular mass, molecular formula, UV, and NMR spectral data. In this chapter, the availability and the suitability of the available databases for dereplication is critically examined. The chapter concludes with examples of dereplication strategy based on available natural product databases.

6.1 Dereplication

One definition of dereplication is the *differentiation of novel metabolites from known compounds in a natural products extract*. The task of dereplication then is one of instituting approaches that achieve this differentiation as quickly and as efficiently as possible. An added bonus is if the process can be achieved on as small a scale as practical. How difficult a task is this? As of June 2011, the Dictionary of Natural Products [1] listed 159,670 unique naturally occurring compounds. In November 2009, the suggested number of natural products in the CAS database [2] was 250,000 while in Reaxys [3] the number was 170,000. While a definitive figure is not possible, it is not unreasonable to assume that there are most probably 175,000 natural products that have been isolated and characterized to date. Thus, the probability that a crude natural products extract will contain new compounds is not high. The history of marine natural products does not extend back as far as that of their terrestrial counterpart, but there are at least 22,000 marine natural products, again offering high odds against the discovery of new compounds. To effectively undertake dereplication of a crude natural product extract, it is first necessary to obtain definitive information about each component of interest in the extract and then compare those data against appropriate chemical and natural product-based databases. The efficiency of this process is very much a function of the quality and accessibility of the relevant data in the available databases. This chapter is focused on surveying the available databases and examining the requirements for dereplication.

6.2 Natural Products Databases

Access to appropriate databases is essential for the efficient study of (marine) natural products whether the aims be the discovery of new compounds, the synthesis of known compounds or analogues, analysis of data relating to taxonomy and distribution of

Table 6.1 A compiled list of databases dealing with natural products

Public domain	Private domain	Commercial
ChemSpider [19]	All Pharma	CAS REGISTRY [2, 4–6]
CSLS [18]	GVK Biosciences NPD [40]	SpecInfo [11]
PubChem [17]	UC UV DB [24]	Reaxys [3]
NMRShift DB [25]	DTU UV DB [22]	ACD/Labs [13]
Naproc-13 [26]	Marine NP DB [41]	NaprAlert [20]
SuperNatural [27]	InterMed UV DB [42]	Dictionary of Natural Products [1]
SDBS [28]	InterMed NMR DB[42]	Dictionary of Marine Natural Products [8]
	Novartis IR DB [43]	AntiBase [10]
	National Centre Plant Metabolism [44]	MarinLit [12]
	CH-NMR-NP [45]	AntiMarin [14]

compounds and source organisms, or studies on bioactivities to name but a few. There are specialized databases of chemical literature, X-ray crystal structure data, NMR spectra, reactions, physicochemical, and bioactivity data. Private domain, free access, and commercial databases abound with an array of different content, coverage and access restrictions. A selection of those considered most relevant is depicted in Table 6.1.

Recent developments in open access and open source resources have generated a rapid increase in the amount of publically available chemical information and the development of sophisticated, free tools to extract these data. Research funded by the US National Institutes of Health (NIH), the Wellcome Trust (British), the Italian National Institute of Health (ISS), the European Research Council, and many other research institutes and universities is required to be made available through a publically accessible repository after acceptance for publication. Freely available data from sources like these provide the core content of other open access databases with specialized search capabilities.

Private Domain databases are those that have been commented on or alluded to in the primary literature. Undoubtedly, there are many more Private Domain databases than those listed in Table 6.1. For dereplication, it is unfortunate that these sources are not generally available as they would be invaluable and would give access to collections of data not otherwise available. The classic examples of this sort of resource are those databases that have been generated by the pharmaceutical companies. Pharmaceutical companies maintain their own extensive libraries of resources, but these are strictly private and inaccessible to other researchers. None of the Private Domain databases will be surveyed here.

There are commercial databases available that contain sufficient substances, data about them, and search capabilities to serve the needs of natural products dereplication, but there is enormous variation in accessibility and fee structure for access to these resources. Most are very expensive, but academic discounts are often available and pricing options are sometimes offered, catering to occasional users or to those who only need access to a portion of the data. The more expensive of these databases have very broad coverage and contain bibliographic entries and substance information of interest to a wide range of users other than natural products chemists or biologists. The less expensive databases are naturally smaller and more specialized, but among these are resources that provide access to the most comprehensive publically available collections of natural products and they are particularly suited to the process of natural products dereplication.

The challenge is to select a database that returns the best quality and quantity of return for the time and money involved in extracting the needed data. For a commentary on relative costs, see Sect. 6.9. The following discussion describes some of the attributes of the more significant of these commercial and open access databases.

6.3 Commercial Databases

6.3.1 CAS REGISTRY and CAPLUS

Undoubtedly, the most comprehensive compilation of information on natural product substances is the Chemical Abstracts Service CAS REGISTRY database in which each substance is identified by a unique CAS Registry Number (RN). Included are chemical structures, trade names, systematic names, synonyms, molecular formula, and calculated and experimental property data. The companion CAPLUS database contains patent and journal article references with abstracts. CAS REGISTRY contains over 63 million organic and inorganic substances, and it is estimated that more than 250,000 of them are natural products [4]. There are a number of options for gaining access to these databases.

6.3.1.1 STN

STN [5] is an online service that provides a single platform for access to over 200 of the most significant science, technology, and patent databases from different suppliers throughout the world, including CAS REGISTRY and CAPLUS. Various means of access include STN Express which permits desktop access to all STN databases. It uses a powerful but complex STN command-language interface that is intended for experienced online searchers. Similar interaction with the selected databases is also provided through a web interface, STN on the Web. For occasional and novice searchers, an easier search interface that uses keywords and Boolean operators (no special command-language is required) is available via web access with STN Easy. An option for one-off search needs is a mediated search to the user's requirements that will be carried out on STN databases by experts who will send the results to the user. This service is available through the FIZ Karlsruhe's search service.

6.3.1.2 SciFinder

The other major access pathway to the CAS REGISTRY and CAplus files is through SciFinder [6], a database produced by the American Chemical Society. SciFinder provides a single, graphical interface to access the CAS REGISTRY and CAplus databases, as well as CASREACT (a reaction database), CHEMCATS (chemical supplier listings), CHEMLIST (regulated chemical information), MARPAT (Markush structures of organic and organometallic molecules from patents), and MEDLINE (the National Library of Medicine database). SciFinder is available on subscription as a web-based and/or client-based system.

Access to the CAS REGISTRY database is mandatory in any MNP investigation that produces a supposedly new compound to ensure that what is being claimed as new is indeed the case. The SciFinder interface is very versatile, permitting searches in various ways to establish the previous occurrence or novelty of a compound, or its similarity to other known compounds. One of the first pieces of experimental information that is often obtained for a compound under investigation is its molecular mass. Somewhat surprisingly, in SciFinder it is not possible to search directly for all substances with a particular mass, but this result can be achieved by first carrying out a substructure search for all compounds containing C and then refining the search based on mass.

The CAS REGISTRY and CAplus files are updated on a daily basis, with data usually being current to within only a few months from publication.

6.3.2 Reaxys

Reaxys [3] contains an extensive repository of experimentally validated data including structures, reactions (including multistep reactions), and physical properties. These data are derived from CrossFire Beilstein, CrossFire Gmelin, and Patent Chemistry Database. While this combined database is primarily designed to meet the needs of synthetic chemists, the easy-to-use web interface is well-suited to the needs of natural products chemists, providing flexible access to information on an estimated 170,000 natural products. The actual number of MNPs in this database is not readily discernable, but is estimated to be quite significant. Reaxys would be particularly valuable to those researchers interested in the synthesis of natural products.

The Reaxys database is updated monthly with data extracted from over 150 journals and from patent offices (US, WO, EP, class C07). The average time from receipt of an article to its inclusion in the database is 6 weeks.

6.3.3 Dictionary of Natural Products and Dictionary of Marine Natural Products

The Chapman & Hall/CRC Press Dictionary of Natural Products (DNP) [1] is a structured database holding information on chemical substances including descriptive and numerical data on chemical, physical, and biological properties of compounds, systematic and common names of compounds, literature references, and structure diagrams and their associated connection tables. DNP is available by annual subscription with desktop data and supporting software on DVD, or access can be obtained through the web-based CHEMnetBASE [7]. Version 20:1 (June 2011) of DNP contained 220,470 entries, of which 159,670 were ascribed to isolated natural products. The additional 60,800 entries were for derivatives of the actual natural products. The Dictionary of Marine Natural Products (DMNP) [8] is a subset of data from DNP based on the biological source of the compounds. DMNP is available as a book with CD-ROM for a desktop version, and also from the web-based CHEMnetBASE. The web version (November 2009) contains 34,685 entries, of which 22,664 are for isolated MNPs, with the balance being for derivatives. The number of isolated MNPs contains a significant number of compounds that were first isolated from terrestrial sources but were subsequently also found in marine organisms. Both the desktop and web-based versions of DNP and DMNP permit flexible searching and reporting options, including substructure searches. Structure results for many compounds do not show stereochemistry in the diagrams. Only the parent compound in a series of related compounds is represented by a stereodiagram, while the related compounds can be viewed as planar diagrams with text descriptions of the variations in configurations. Linear peptides are shown as sequences rather than as diagrams. A recent reviewer of DMNP found issue with a number of the offered features of the electronic dictionary [9]. Updated DVD and CHEMnetBASE versions of DNP are released every 6 months, while the CHEMnetBASE version of DMNP is updated annually. Each release has data current to within 6–12 months of publication.

6.3.4 AntiBase

AntiBase 2011 [10] is a comprehensive desktop database of 36,000 natural compounds from microorganisms and higher fungi. AntiBase includes descriptive data (molecular formula and mass, elemental composition, CAS registry number), physicochemical data (melting point, optical rotation), some spectroscopic data (UV, ^{13}C- and ^{1}H-NMR, IR, and mass spectra), biological data (pharmacological activity, toxicity), information on origin and isolation, and a summary of literature sources. A feature of AntiBase is the use of predicted ^{13}C-NMR spectra (specinfo [11]) for those compounds where no measured spectra are available. This database is becoming increasingly important for MNP investigations as more MNPs are discovered from microorganisms where the overlap between “marine” microorganisms and “terrestrial” microorganisms can be difficult to determine. Having knowledge of the chemistry of terrestrial microorganisms is therefore highly desirable. AntiBase is updated annually with data current to within only a few months of publication.

6.3.5 MarinLit

With the exception of CAplus, all of the databases described above are compound-centric. MarinLit [12] is a desktop database comprising records relating to publications covering all aspects of MNP research. Thus, not all entries contain information on newly isolated MNP structures, but they may cover aspects of synthesis, biosynthesis, ecological studies, bioactivity investigations, reisolation of known compounds, and also reviews. MarinLit currently has 24,000 records of publications, of which 8,500 describe 22,000 structures for compounds isolated for the first time from marine sources. All records contain the usual bibliographic information, an extensive list of keywords, and where appropriate, taxonomy, structures, MW, formulae, UV data, and calculated (ACD/Labs [13]) ^{1}H and ^{13}C NMR chemical shifts. Very flexible searching and reporting options are available for combinations from all of these data fields including substructure searching. A unique feature in MarinLit (and AntiMarin – see below) is the inclusion of searchable data fields containing the numbers of each structural feature (^{1}H-SF) that can be determined from inspection of the ^{1}H-NMR spectrum of a compound. These features include the numbers of methyl groups of various types – singlets, doublets, triplets, or -*O*Me, -*N*Me, -*S*Me, etc., types of substituted benzene rings, and numbers of sp^{2}-H, sp^{3}-CH or sp^{3}-CH$_{2}$ groups to name a few. The value of this feature for MNP dereplication and discovery investigations will be described later in this chapter. MarinLit updates are released twice each year with the data being current to within 1–2 months of publication.

6.3.6 AntiMarin

AntiMarin [14] is available to current subscribers to both AntiBase and MarinLit. It is a compound-centric desktop database containing data for 53,000 compounds from AntiBase and MarinLit. The compound data from each database is enhanced by the inclusion of searchable data fields containing the numbers of each structural feature that can be determined from inspection of the ^{1}H-NMR spectrum of a compound, as described above for MarinLit. This combined database with the structural features data included provides a valuable tool for the process of dereplication, as described later in this chapter. AntiMarin is updated annually with data current to within a few months of publication.

6.3.7 Spectroscopic Databases

While the previously described databases contain some spectroscopic data for compounds, or at least reference to the source of experimental spectroscopic data for a compound, there are other databases dedicated to the cataloging and/or calculation of spectroscopic properties. Access to these data can be particularly helpful in the investigation of MNPs, either to establish that the experimental data obtained in an investigation is the same as that previously found for a proposed structure, or to determine if the observed data for a new proposed structure is reasonable. The following descriptions do not include packages that attempt to generate a structure from experimentally obtained data.

6.3.7.1 SpecInfo

SpecInfo [11] is a spectroscopic database whose primary aim is to assist with spectral interpretation and structure elucidation. These functions are supported by facilities for searching the database for compounds with NMR and/or IR spectra, or fragments of spectra, matching the experimental data. Additionally, compounds can be searched for using structures or substructures or other structurally related information such as CAS numbers. NMR spectrum prediction for a proposed structure is also an integral part of SpecInfo. These capabilities are supported by a knowledge base of 359,000 ^{13}C NMR spectra, 130,000 ^{1}H NMR spectra, 90,000 heteroatom (^{15}N, ^{17}O, ^{19}F, ^{11}B and ^{31}P) NMR spectra, 139,000 mass spectra, and 21,000 IR spectra.

6.3.7.2 ACD/Labs

ACD/Labs [13] provide a collection of software packages directed principally at the handling and processing of experimental data, mostly spectroscopic. For the natural products chemists, the packages of most interest would be the HNMR and CNMR Predictors. These permit the calculation of ^{1}H and ^{13}C NMR spectra from user-inputted structures. These Predictors utilize algorithms based on more than 1.7 million assigned ^{1}H chemical shifts from more than 210,000 chemical structures and 2.5 million assigned ^{13}C chemical shifts from about 200,000 chemical structures. Use of these Predictors can be very helpful for comparing the calculated spectra for a proposed structure with the experimental data to assess the feasibility of a proposed structure. Further tools are available in ACD/NMR Workbook for more direct comparisons of calculated 2D spectra with the experimental spectra, again assisting with the verification of a proposed structure. Of particular relevance to MNP chemists are the internal databases in the Predictor packages. These contain the published chemical shift data for ~240,000 compounds. These data are only entered into the internal databases after a rigorous analysis of the data to establish that the assignments as published are consistent with those arrived at by calculation. Presently, over half of the marine natural products have their data included in the internal databases, and these data are being added to on a regular basis so that the proportion of MNPs contained in the internal databases will eventually be much higher. Currently, the ACD/Labs-calculated ^{1}H and ^{13}C NMR chemical shift data for all MNPs are accessible from within MarinLit, as described earlier. Additional useful data in the internal databases are the original references, solvents, frequency, NMR techniques, molecular formulae, molecular masses, IUPAC names, and trivial names, all of which can be searched, viewed, and printed.

6.4 Free Access Databases

There are now over 50 databases with free access to chemical structures from various sources [15, 16]. However, of more use to natural products chemists are those databases containing compound data collected from a wide range of other open access or proprietary databases. In general, these combined databases contain the chemical structures and a limited amount of other associated data but do not refer to the source of the compounds. They do, however, provide links to the source database from which the data were compiled, and thus allowing the user to make their own arrangements for access to the complete information relating to a compound. A particularly valuable feature of these combined databases is that they allow the user to determine if a structure may have been previously characterized, although this will not be a complete substitute for verifying novelty of a compound as necessary through the use of the CAS REGISTRY. In general, these databases do not provide comprehensive classification of compounds that might be natural products. It is not possible to describe all of the databases that are available, and the following descriptions are for those that are likely to be the most generally useful for natural products studies.

6.4.1 PubChem

PubChem [17] includes substance information, compound structures, and bioactivity data in three primary databases: PC Substance, PC Compound, and PCBioAssay, respectively, with data collated from over 80 other databases. PC Compound contains more than 25 million unique structures.

6.4.2 CSLS

The Chemical Structure Lookup Service (CSLS) [18] can be regarded as an address book for chemical structures. It has two major modes of operation: The first mode permits the submission of one or more chemical structures in the form of an SD file, as SMILES strings, or in one of more than 20 other molecular structure formats that CSLS understands. The service will determine whether the submitted structures are present in any of the databases that are currently indexed in CSLS. The second mode allows the submission of a document, from which CSLS will attempt to extract all possible chemical information that this document might contain – InChI string, InChIKey, SMILES string, molecular formula, or NCI/CADD Structure Identifiers (uuuuu, FICuS, or FICTS). It then conducts a search with these extracted chemical data. There are 74 million entries in CSLS collated from about 100 databases, representing 46 million unique structures. In the classification section of CSLS, there is a check box for natural products. However, this refers to about 124,000 entries from the NCI Frederick NP database, of which only about one third are actual natural products. An additional 8,000 natural product structures from the CHMIS-C database are included in this section.

6.4.3 ChemSpider

ChemSpider [19] is a compound-centric search engine and database, now being developed under the auspices of the Royal Society of Chemistry, which is aggregating and indexing chemical structures and their associated information into a single searchable repository. This database aims to capture and manage chemical structures from online resources, from commercial databases, and from users of the ChemSpider platform who have the ability to submit their own data. Users can access the open access data immediately and where possible, the ChemSpider search engine also provides links to commercial resources that contain information matching the users' query. Many additional properties have been added to each of the chemical structures thus enhancing the value of the collection. These include spectral data, links to publications, reaction synthesis details, and various experimental properties. For MNP researchers, perhaps the greatest value will be to determine if there is any information available about a compound of interest. Access to ChemSpider is without charge. Of particular value is the ability to search by substructure, a feature that is not available in CSLS. Currently (2011), there are in excess of 26 million unique entries in ChemSpider from over 300 data sources. There are numerous cases of the same natural product in the database with various levels of partial to complete stereochemistry as provided by a number of the depositors. Members of the ChemSpider team are focused on curating the structure collection. The information in ChemSpider is updated daily as a result of new compound depositions and curation activities, but the currency and accuracy of the data is only as good as that of the databases from which the data is sourced.

6.5 Approaches to Dereplication

In a dereplication exercise, a minimal set of must-know data would include the molecular masses (and molecular formulas) of the components of the mixture, relevant UV data and, if possible, ^{1}H NMR spectra. The taxonomy of the organism, although very useful, is not an absolute requirement.

The molecular mass and UV data can be generated from an LC/MS analysis of an aliquot of the crude extract using Diode Array Detection (DAD) in combination with electrospray mass spectrometric analysis (ESMS). Under high-resolution conditions (LC/HRESMS), the individual molecular formulas of the components can be obtained. With access to a CapNMR probe or small-tube cryoprobe, it is now possible to obtain a good ^{1}H NMR spectrum of individual components from the same LC run (see Sect. 6.8). The ideal situation would be to use this minimal set of data, perhaps in combination with taxonomic data if it is available, and make a definitive identification of the compound as either new or known. Of course, there are other ways that this minimal data set could be generated, for example, by chromatography on the crude extract followed by MS or HRMS, ^{1}H NMR, and UV measurements on the isolated individual components. Taxonomic identification of marine organisms can be fraught with difficulties, but this knowledge undoubtedly can be of assistance.

6.5.1 Selection of Database

Not all of the natural product databases suggested (Table 6.1) can give definitive answers using part, or even all of this suggested data set. In Table 6.2, a "filtered" list of databases has been provided listing the attributes and the data that can be extracted readily.

In Table 6.2, the databases are arranged from the largest (CAS REGISTRY) to the smallest (MarinLit). The smaller databases, from NaprAlert [20] downward, are the dedicated natural product databases. The number of natural products in each database has been listed. In the larger databases such as SciFinder and Reaxys this is an estimate only. With the exception of NaprAlert all of the databases are kept current, or within a few months of current. NaprAlert's coverage of MNPs since 2004 has been sporadic only and for this reason has not been covered in this chapter. Within the range of databases in Table 6.2, molecular formula-based searches are possible and, with the exception of CSLS, it is possible to search on a molecular mass range, although doing so within SciFinder is not obvious as it is first necessary to have generated a subset from SciFinder that contains all compounds containing C and then to initiate a search using the molecular mass range of interest as a filter. All of the databases except NaprAlert are capable of carrying out substructure searching. This is a particularly helpful feature for using recognizable fragments in searches. These fragments can arise from interpretation of NMR and/or MS data. As well, taxonomic and biological activity data can be searched in most of the databases listed. The distinction between the utility of the various databases comes when the availability of actual UV and NMR data is considered. For UV data (λ and ε), the two Dictionaries and MarinLit contain this data with partial coverage included in NaprAlert, AntiMarin and AntiBase. For NMR data within this group of databases, δ_C values are searchable within MarinLit. The only databases available that can provide spectral data are MarinLit and AntiBase. Through an arrangement with ACD/Labs, *calculated* ^{1}H, ^{13}C chemical shift data and HSQC/DEPT spectra are accessible in MarinLit. This is a facility that is particularly useful for comparing actual data from a potential new compound against data that have been generated for known compounds. The last NMR feature listed in this figure is ^{1}H NMR **S**tructural **F**eatures (^{1}H-SF). This is a unique aspect for searching ^{1}H NMR data for matching features and is only available using the MarinLit or AntiMarin databases. As noted earlier, these two databases include searchable data fields containing the actual numbers of each

Table 6.2 A “filtered” list of databases that are of potential use for the dereplication of (marine) natural product extracts

	Number of compounds[a]									NMR data[b]			
Database	Total	Natural products	Current up to	MW	MF	UV[c] λ	SSS[d]	Tax.[e]	Biol.[f]	δ	Spectra	^{1}H-SF	HSQC/DEPT
CAS REGISTRY	6.3×10^7	~250,000	Current	+[g]	+	–	+	+	+	–	–	–	–
CSLS	4.6×10^7	?	Current	–	+	–	+	–	?	–	–	–	–
ChemSpider	2.6×10^7	?	Current	+	+	–	+	–	–	–	–	–	–
PubChem	3.7×10^7	?	Current	+	+	–	+	–	+	–	–	–	–
Reaxys	$>10^7$	170,000	Current	+	+	–	+	+	?	–	–	–	–
NaprAlert		>150,261	~2004	+	+	+[h]	–	+	+	–	–	–	–
Dictionary of NaturalProducts		159,670	2011	+	+	+	+	+	+	–	–	–	–
AntiMarin		43,000	2010	+	+	+[h]	+	+	+[h]	–	–	+	–
AntiBase		36,000	2010	+	+	+[h]	+	+	+	+[h]	–	–	–
Dictionary of Marine Natural Products		22,691	2006	+	+	+	+	+	+	–	–	–	–
MarinLit		22,000	2011	+	+	+	+	+	+[h]	+	+	+	+

[a] Where possible an estimate is given for the number of natural products in the database
[b] Four options for NMR data: the δ values (calculated or actual), spectra, ^{1}H NMR structural features (^{1}H-SF) or calculated HSQC/DEPT spectra
[c] Actual λ(ε) values for UV data as opposed to a reference to the data
[d] Sub-structure searching capability
[e] Taxonomic data
[f] Biological activity data
[g] In the current version of SciFinder the extraction of molecular mass data is not straightforward
[h] Partial data only

structural feature that can be deduced from the ^{1}H NMR spectrum of a compound. For example, the numbers of methyl groups of various types – singlets, doublets, triplets, or -*O*Me, -*N*Me, -*S*Me, etc. This unique feature, in combination with mass and perhaps UV data, is very effective in discriminating between alternative structures in the dereplication process, as will be described in Sect. 6.6.

6.5.2 Dereplication Based on Molecular Mass/Molecular Formula

An early step in dereplication is to obtain the MS of compounds isolated by chromatography of the crude extract or by running an LC/MS experiment on the crude sample. Depending on the resolution and/or mode of the MS or LC/MS, two outcomes are possible. If the MS, typically ESMS, has been run under low resolution conditions (<1:5,000) then unit mass differentiation is possible. That is the distinction between say *m*/*z* 490 and 491. Under higher resolution conditions, the actual molecular formula can be obtained. For example, $C_{30}H_{50}O_5$ which has *m*/*z* = 490.3658. Either or both of these outcomes can be searched in databases. The results of such a search are shown in Table 6.3. Even for the more specialized databases, the number of “hits” recorded is often too great even when searching for a molecular formula. Sometimes molecular mass data is all that is available, but that alone is not a good discriminator. Ideally, less than ten hits is an acceptable number.

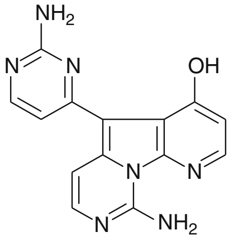
If a new compound that has an unusual or a unique molecular formula, obtaining the molecular formula may be all that is necessary to identify it as a new compound. That was the case with variolin B, a bioactive alkaloid isolated from the Antarctic sponge *Kirkpatrickia varialosa* [21]. Variolin B (**1**) had a molecular formula of $C_{14}H_{11}N_7O$. At the time a search of the specialist natural product databases gave zero hits thus establishing that this was a new compound. Nearly 20 years later,

Table 6.3 Search of relevant databases from both free access and commercial sources for molecular mass and molecular formula data

		$C_{30}H_{50}O_5$
Database	*m*/*z* 490–491	(*m*/*z* 490.3658)
ChemSpider	58,938	173
CAS REGISTRY	171,904	2,366
CSLS	na[a]	250 (81NP’s)
Dictionary of Natural Products	653	292
Dictionary of Marine Natural Products	94	44
MarinLit	78	35
AntiMarin	118	43

[a] It is not possible, at this stage of development, to search CSLS for molecular mass information. Molecular formula searches can be carried out and 250 matches found with 81 of these in the Natural Product section of the database

MarinLit still records variolin B as the only compound with that molecular formula while in SciFinder there are 79 compounds recorded with that molecular formula. However, it is usually necessary to have more than just molecular mass/molecular formula data for the dereplication process.



1

When dealing with natural product extracts, there can be problems and uncertainty in obtaining reliable molecular mass/molecular formula data. This could arise because:

- the mass spectrometry sample is only one or two fractionation steps removed from the crude extract, and there are multiple candidates for the supposed molecular ion from impurities;
- such impurities could dominate the ionization giving misleading results;
- there can be ionization suppression problems from traces of TFA (often used as a polar modifier in HPLC);
- of ready fragmentation, even under ESMS conditions;
- of formation of adduct ions (MNa^+, $MNH_4{}^+$, etc.).

Some of these problems have been very nicely addressed by a group at the Danish Technical University group who, in the area of mycotoxins and fungal metabolites, compiled a database of 474 compounds using standardized HPLC/UV/MS methodology [22] (see Table 6.1).

6.5.3 Dereplication Based on UV Data

UV profiles or maxima are readily acquired using a Diode Array Detector (DAD) as part of the LC or LC/MS examination of a crude extract. That the data are semiquantitative at best is not relevant. What is important are the actual profiles, or the maxima (λ_{max}), as the chromophores that lead to these spectral properties are distinctive and can be searched for. The UV spectra and λ_{max} are indicative of a *chromophore* within a structure, not necessarily the structure itself and therefore offers clues as to potential substructures that can be searched for, for example, the 1,2,3,5-tetrasubstituted aromatic system characteristically found in compounds of polyketide origin. Compounds containing this chromophore have a characteristic UV profile with

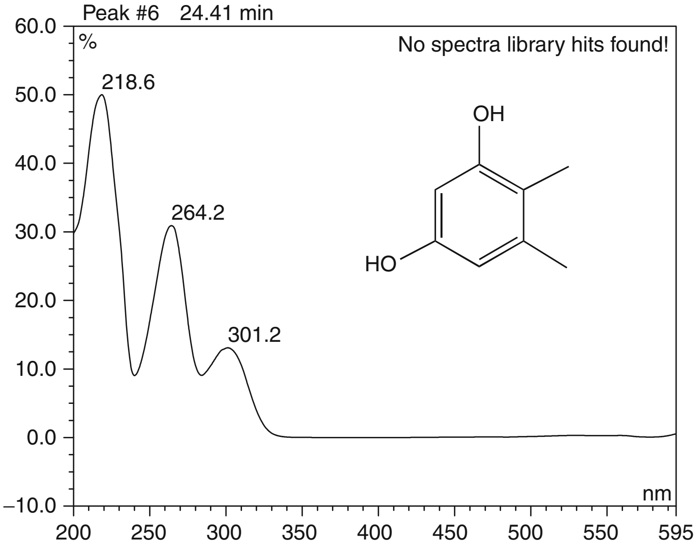


Fig. 6.1 The characteristic UV profile of a 1,2,3,5-tetrasubstituted aromatic system

λ_{max} at around 220, 265, and 300 nm (see Fig. 6.1) that is easily recognizable. Using UV data with mass data is a powerful and cheap method of dereplication. The one major drawback to this approach is that not all compounds contain UV chromophores.

6.5.3.1 UV λ_{max} Data

Among the natural product databases, a number contains searchable UV λ_{max} data, though in some cases this is partial coverage only (see Table 6.2). An example of this approach to dereplication is work that was carried out on a bioactive extract obtained from the deep-water sponge *Lamellomorpha strongylata* [23]. All of the bioactivity associated with this extract eluted in the early fractions from an LH-20 column (higher molecular mass compounds) and the two components in these early fractions each had identical and characteristic chromophores (see Fig. 6.2).

A search was made in MarinLit using the following search profile: Phylum = Porifera; Mass range 750–2,000; UV = 340 (see Fig. 6.3). The database returned 16 matches from 22,000 possible compounds. Within MarinLit it is possible to select a second UV maxima and refine the search. This was done using the second maxima at 226 nm and resulted in just nine compounds that matched. All but one of these compounds were calyculin derivatives. Based

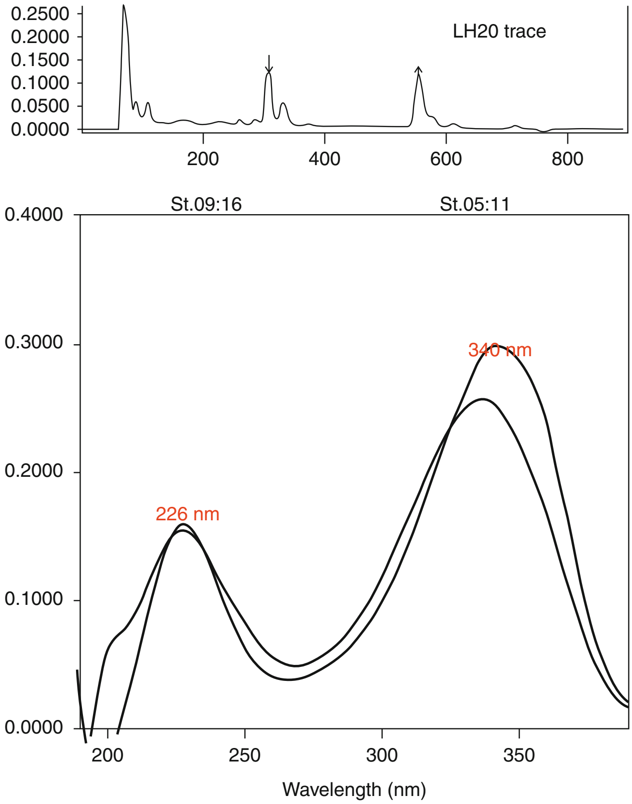
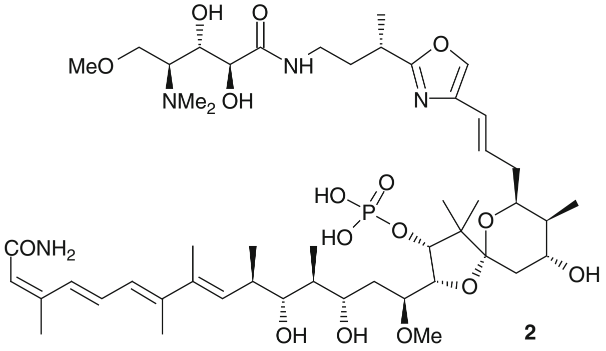
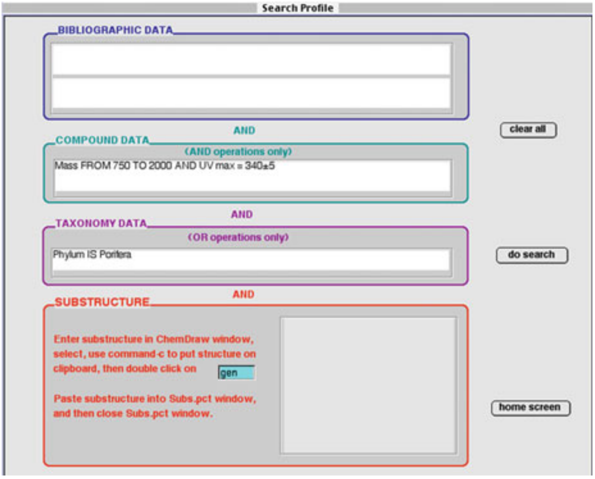


Fig. 6.2 HPLC trace (detection at 340 nm) and extracted UV data for the peaks at 320 and 550 s from an LH-20 column (Fraction 1)

on these data, the two bioactive compounds from *Lamellomorpha strongylata* were rapidly identified as calyculin A and a new, but related compound, calyculinamide A (**2**).



2



Mass FROM 750 TO 2000 AND UV max = 340±5

Phylum IS Porifera

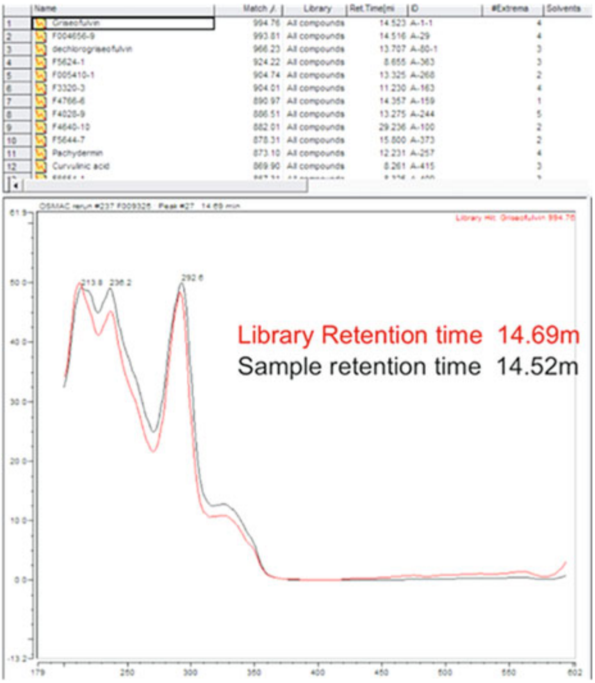
Enter substructure in ChemDraw window, select, use command-c to put structure on clipboard, then double click on gen

Paste substructure into Subs.pct window, and then close Subs.pct window.

Fig. 6.3 Search profile in MarinLit

6.5.3.2 UV Spectra

There is considerably more information contained in the UV spectrum than just the λ_{max} data. Matching spectra is a superior and more definitive approach to the use of UV spectra in the dereplication exercise. Such approaches have been commented on in the primary literature, and it can be assumed that all pharmas carrying out natural product research and quite a number of other natural product laboratories have in-house UV spectral databases. Unfortunately, no UV spectral databases that also contain other information essential to the dereplication process are available. Apart from privacy and IP issues, a major difficulty is the platform to be used for comparing the spectra. Most modern HPLCs have the necessary software for capturing and comparing UV acquired spectral data, but not for comparing that data with spectra acquired on other HPLCs. For the past 10 years in the Marine Group at the University of Canterbury, all UV data from all extracts is archived in a searchable library (All Compounds). Once the identity of compound is established, the UV profile is added to a second library (Known Compounds). As all HPLC runs have been carried out using the same solvents, column manufacturer, and gradient profile the retention times as well as the UV profiles of unknowns can be run against the library and, frequently, unknowns can be identified by UV/retention time correlations only [24]. An example illustrates just how useful this approach can be for compounds with distinctive UV chromophores. By this approach, griseofulvin was identified as a metabolite from a marine fungus (see Fig. 6.4). Note the close match between the retention times of the unknown and the reference sample as well as the comparable UV spectra. The library provides a score, out of 1,000, for the closeness of the spectral match – the griseofulvin score was 994. As appropriate, the retention time window can be widened or eliminated altogether in order to match against the UV spectrum only.



	Name	Match /.	Library	Ret.Time[mi	ID	#Extrema	Solvents
1	Griseofulvin	994.76	All compounds	14.523	A-1-1	4	
2	F004656-9	993.81	All compounds	14.516	A-29	4	
3	dechlorogriseofulvin	966.23	All compounds	13.707	A-80-1	3	
4	F5624-1	924.22	All compounds	8.655	A-363	3	
5	F005410-1	904.74	All compounds	13.325	A-268	2	
6	F3320-3	904.01	All compounds	11.230	A-163	4	
7	F4766-6	890.97	All compounds	14.357	A-159	1	
8	F4028-9	886.51	All compounds	13.275	A-244	5	
9	F4640-10	882.01	All compounds	29.236	A-100	2	
10	F5644-7	878.31	All compounds	15.800	A-373	2	
11	Pachydermin	873.10	All compounds	12.231	A-257	4	
12	Curvulinic acid	869.90	All compounds	8.261	A-415	3	

OSMAC rerun #237 F009326 Peak #27 14.69 min

Library Hit: Griseofulvin 994.76

Library Retention time 14.69m

Sample retention time 14.52m

Fig. 6.4 DAD UV spectrum of griseofulvin compared with that stored in the known compounds library

6.5.4 NMR-Based Approaches to Dereplication

^{1}H NMR spectra are rich in structural information, which in combination with 2D homo- and heteronuclear experiments, molecular mass, molecular formula, and UV data lead to structural assignments. Inherently, there are two drawbacks to the routine use of NMR techniques for dereplication purposes where a rapid answer to the question of novelty is required. The first of these is sensitivity. Of the spectroscopic techniques routinely used in organic structure determination, NMR spectroscopy is by far and away the least sensitive. The limit of detection for routine mass spectrometers is in the 1–10 pg range (10^{-12} g), that for UV spectroscopy 100–1,000 pg, while for most routine NMR spectrometers, 500–1,000 μg are required for ^{1}H NMR measurements and about five times that for ^{13}C NMR spectroscopy. In recent years, the limit of detection for ^{1}H NMR has dropped to 1 μg using specialist probes (see Sect. 6.8). This is still 10^{6} times less sensitive than mass spectrometry. The implications are that it is possible to obtain excellent LC/MS and associated UV data by the analysis of micrograms of crude extract, but to get NMR data on the components of a crude extract, it is necessary to process mgs or even grams of crude extract.

The second drawback is that of complexity. In order to derive substructures for database searching, interpretation of the 1D and 2D NMR data is required.

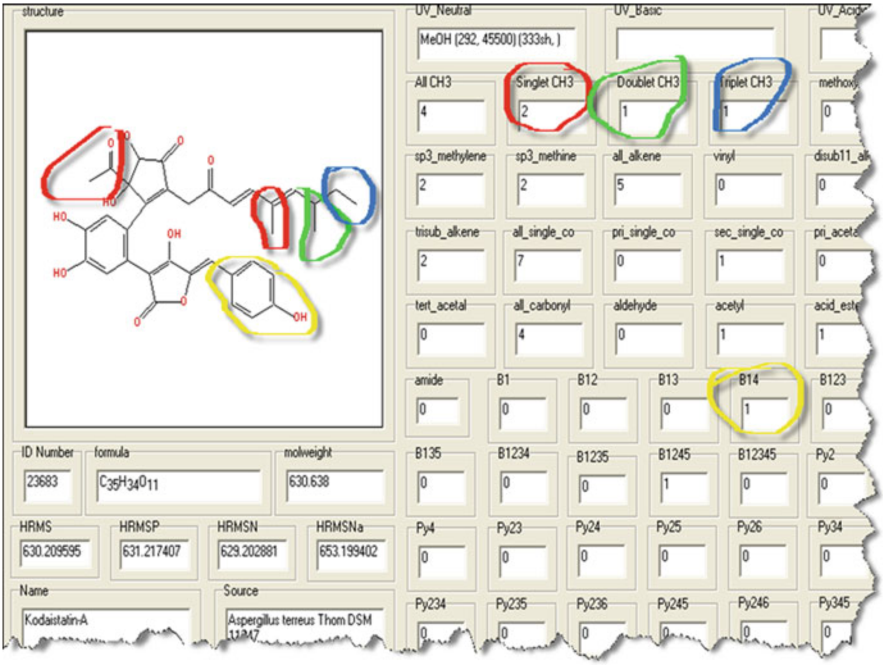
Depending on the complexity of the molecule this can sometimes be a complex process, but the derived substructures can then be searched for in any of the natural product databases (Table 6.2). There are also spectral databases available, such as ACD/Labs and SpecInfo, which can be searched to find a match based on molecular formula comparisons. And in the Public Domain arena (Table 6.1), there are other spectral sources that can be searched. These are the likes of NMRShift DB, Naproc-13, SuperNatural and SDBS [25–28]. From the natural product databases listed in Table 6.1, only AntiBase and MarinLit have searchable NMR data. MarinLit, in an arrangement with ACD/Labs, can provide calculated ^{1}H and ^{13}C chemical shift data for any of the individual compounds in the database along with ^{13}C and HSQC/DEPT spectra based on the ACD predictors (see Sect. 6.3.7.2), allowing the ready checking of actual NMR data against that stored for individual compounds in MarinLit identified using other parameters. MarinLit also allows for the direct searching of the database for individual or a series of ^{13}C chemical shifts. AntiBase provides partial lists of actual or calculated ^{1}H and/or ^{13}C NMR data which are also searchable.

AntiMarin is a combination of parts of the MarinLit and AntiBase databases. Both this combined database and MarinLit have a search capability that is not readily available in any other database. This is the capability of searching for the actual numbers of functional groups contained within a molecule. Certain features in a ^{1}H NMR spectrum are immediately obvious and do not need any interpretation to know what they are. The number and types of methyl groups in a molecule would be a good example. Within these two databases, the number and type of methyl group, alkenes, carbinol protons, acetal, formyl, acetyl, amide, imine, aromatic substitution patterns, sp^{3} methines, and methylenes and sp^{2} H have been extracted and placed in searchable fields. A simple inspection of a ^{1}H NMR spectrum and integrals immediately allows the identification of many of the classes of functional group listed above, but *with no consideration* of any relative connectivity. Entry of the relevant numbers for each functional group, along with other relevant data such as taxonomy, molecular mass, molecular formula, provides a very effective method for the dereplication of natural products extracts as the ^{1}H structural features (^{1}H-SF) aspect built into AntiMarin and MarinLit is very discriminating.

6.5.4.1 Why ^{1}H-SF is Discriminatory

^{1}H structural features (^{1}H-SF) allows the examination of combinations of structural features in a molecule. The probability of compounds having identical combinations of ^{1}H NMR features is low, and if these data are also taken in combination with molecular mass, molecular formula and UV data unique search patterns are generated which can quickly establish the novelty of an isolated compound. The page for each compound in AntiMarin displays the relevant UV, MS, and ^{1}H-SF features for that compound. This is displayed in Fig. 6.5 which highlights features such as the Me groups and the 1,4-disubstituted benzene. The search data are entered via a simple graphical interface comparable to that shown in the figure.

Two simple examples will serve to demonstrate this selectivity. In the first instance, the simple act of counting the number of methyl groups in a compound

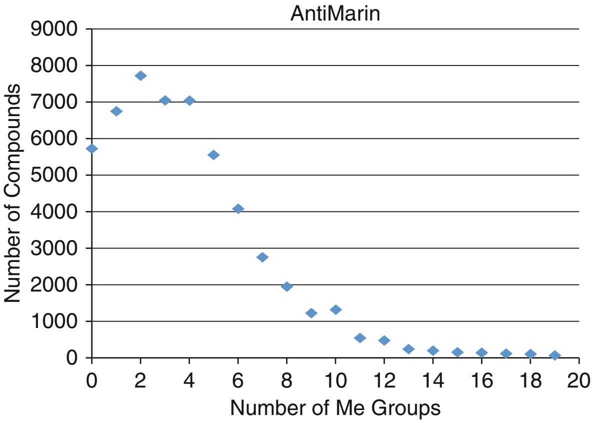


UV_Neutral		UV_Basic		UV_Acid	
MeOH (292, 45500) (333sh,)					
All CH3	Singlet CH3	Doublet CH3	Triplet CH3	methoxy	
4	2	1	1	0	
sp3_methylene	sp3_methine	all_alkene	vinyl	disub11_al	
2	2	5	0	0	
trisub_alkene	all_single_co	pri_single_co	sec_single_co	pri_aceta	
2	7	0	1	0	
tert_acetal	all_carbonyl	aldehyde	acetyl	acid_est	
0	4	0	1	1	
amide	B1	B12	B13	B14	B123
0	0	0	0	1	0
B135	B1234	B1235	B1245	B12345	Py2
0	0	0	1	0	0
Py4	Py23	Py24	Py25	Py26	Py34
0	0	0	0	0	0
Py234	Py235	Py236	Py245	Py246	Py345
0	0	0	0	0	0

ID Number	formula		molweight
23683	$C_{35}H_{34}O_{11}$		630.638
HRMS	HRMSP	HRMSN	HRMSNa
630.209595	631.217407	629.202881	653.199402
Name		Source	
Kodaistatin-A		Aspergillus terreus Thom DSM 11247	

Fig. 6.5 ANTIMARIN showing a range of the data in searchable fields

Fig. 6.6 Distribution of compounds containing methyl groups in the ANTIMARIN database



is informative. Figure 6.6 shows the distribution of the ~52,000 compounds containing methyl groups in the ANTIMARIN database. Of particular relevance is the large number of compounds (~6,000) with zero methyl groups. The second example focuses on the types of methyl group recognized in the MARINLIT and ANTIMARIN databases. There are eight types in all (singlet Me, doublet Me, triplet Me, -*O*Me, -*S*Me, -*N*Me, vinyl Me, and acetyl Me). For compounds that contain any two methyl

Table 6.4 The distribution of combinations of any two methyl groups in the ANTIMARIN database

	Singlet	Doublet	Triplet	Vinyl	Acetyl	-*O*Me	-*N*Me	-*S*Me
Singlet	1,000							
Doublet	268	1011						
Triplet	153	589	813					
Vinyl	279	385	151	400				
Acetyl	72	168	244	77	93			
*O*Me	509	435	401	217	148	490		
*N*Me	78	77	50	30	18	132	223	
*S*Me	0	4	2	6	5	17	16	91

4 × Me groups (any type)	7,674 hits, or
1 × formyl group	2,578 hits, or
1 × 1,4-disubstituted benzene	2,201 hits, or
1 × 1,2,4-trisubstituted benzene	2,425 hits, or
1 × 1,2,3,5-tetrasubstituted benzene	3,090 hits

But

4 × Me (any type) + 1 x formyl	358 hits, or
2 × Me(s) + 2 × Me(d) + 1 × formyl	43 hits, or
2 × Me(s) + 2 × Me(d) + 1 × 1,4-disubstituted benzene	13 hits, or
2 × Me(s) + 2 × Me(d) + 1 × 1,2,4-disubstituted benzene	8 hits, or
2 × Me(s) + 2 × Me(d) + 1 × 1,2,3,5-tetrasubstituted benzene	3 hits

Fig. 6.7 Results from searching on easily recognized functional groups or combinations

groups, 8,385 in the database, there are 36 possible combinations. The distribution from searching all 36 combinations is shown in Table 6.4 and illustrates just how discriminating this approach to dereplication is with all possible combinations of two methyl groups out of eight types being populated to one level or another. That was just using methyl groups as the discriminator. If other easily recognized groups such as formyl, 1,4-disubstituted-, 1,2,4-trisubstituted-, and 1,2,3,5-tetrasubstituted benzenes are now added into the mix, the level of discrimination rises. This is shown in Fig. 6.7. Quite large numbers of possible hits are obtained by searching on individual groups, but by looking at combinations, the number of possible hits is rapidly narrowed (Table 6.4 and Fig. 6.7).

Take the example of the search based on a ^{1}H NMR spectrum that contained two singlet and two doublet methyl groups and a 1,2,3,5-tetrasubstituted benzene. Using the ^{1}H NMR data alone, as detailed in Fig. 6.7, the search was narrowed to just three possibilities from 53,000 compounds (see Fig. 6.8) and if low resolution mass data was then added (ESMS: MH^{+} m/z = 321), the unknown could be tentatively identified as debromohamigeran E (that was originally isolated from the sponge *Hamigera taragensis*) [29]. To confirm that assignment, the original literature would now be consulted and direct comparisons made with the NMR and other relevant spectral data.

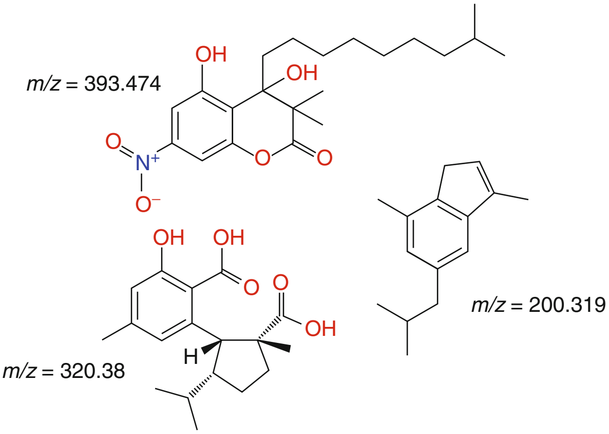


Fig. 6.8 Candidates from search on 2xMe(d) + 2xMe(s) + 1,2,3,5-tetrasubstituted benzene

6.6 Examples of Dereplication

Very seldom is it possible to dereplicate a crude extract without accessing several pieces of information about the components in the extract. Most often these data would be combinations of the source taxonomy, molecular mass/molecular formula, UV, and NMR data. With access to appropriate natural product databases, it is then possible to verify the novelty or not of the components of the extract. In the sections that follow, several worked examples will show how this can be achieved.

6.6.1 Compound Isolated from a Cnidarian

Several compounds were isolated from the crude extract of a cnidarian, possibly a *Minabea* sp. The compound of interest had a low resolution molecular mass of 470, had a UV λ_{max} at 240 nm, and a ^{1}H NMR spectrum which contained a number of easily recognizable features (see Fig. 6.9).

6.6.1.1 Taxonomy Approach

A search in MarinLit using Phylum = Cnidaria gave a total of 2,807 articles containing 4,491 compounds. This clearly is not sufficiently discriminating, but if the cnidarian was indeed a *Minabea* sp., the search is narrowed down to just five articles and 21 known compounds. When the mass data of the compound of interest

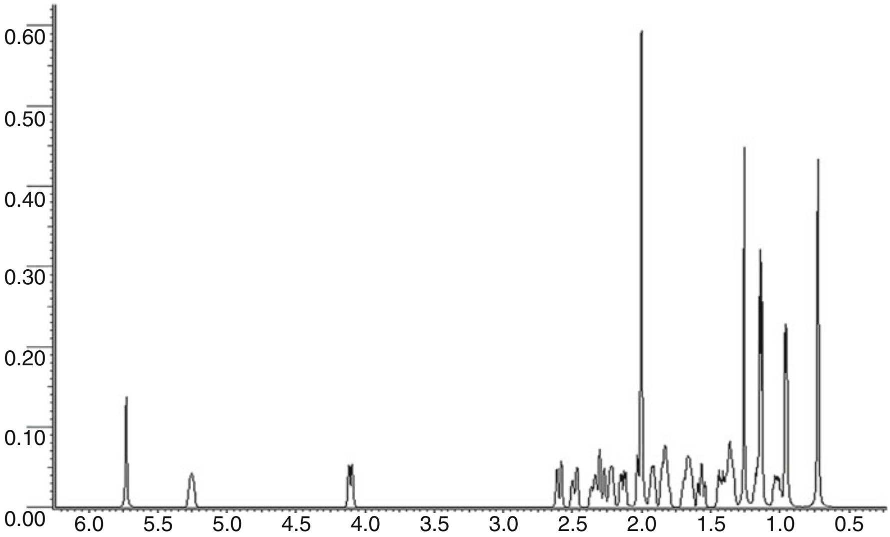
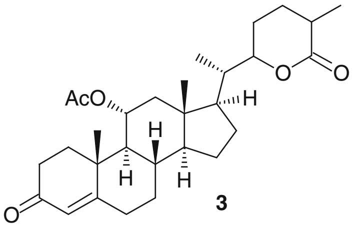


Fig. 6.9 ^{1}NMR spectrum of compound isolated from a cnidarian

was searched against the database, four compounds matched and all had the molecular formula $C_{29}H_{42}O_5$. If the UV data, indicative of an αβ-unsaturated ketone, is added into the profile, only one compound, minabeolide-8 (3), matched. This process would have effectively dereplicated the compound of interest in that extract.



3

6.6.1.2 Alternative Approaches

If the taxonomic data that in combination with the mass and UV data allowed assignment of structure had not been available, what alternative approaches could have been made? There are several possibilities. Examination of the ^{1}H NMR data suggests that the compound contains five methyl groups of which three are singlets and two are doublets. Of the singlets, one of these (δ = 2.05) could readily be assigned as an -*O*Ac. Using search profiles in MarinLit gives the following results:

Cnidaria	2,807 articles/4,491 compounds
Cnidaria + $\lambda_{max} = 240$ nm	198 articles/355 compounds
Cnidaria + UV + *m*/*z* = 470–471	5 articles/5 compounds
Cnidaria + UV + *m*/*z* = 470–471 + 5 × Me groups	1 article/1 compound

giving the same answer as before.

An alternative search profile based initially on ^{1}H NMR data could be

5 × Me groups	1,774 articles/3,166 compounds
5 × Me groups + 3x singlet/2x doublet	264 articles/424 compounds
5 × Me groups + 3x singlet/2x doublet + 1x-*O*Ac	64 articles/77 compounds

If the source phylum is now entered, the number of articles is 42 articles/33 compounds and, finally with the mass data, the numbers drop to 1 article/1 compound.

What if only NMR data were used? There are other resonances that can be used from the ^{1}H NMR spectrum. For instance, the 1-H resonances at δ 4.2, 5.25, and 5.75 can readily be assigned as 2x-CH-O- and 1x > C = CH-. Using these data produces the following:

5 × Me groups + 3x singlet/2x doublet + 1x-*O*Ac	64 articles/77 compounds
5 × Me groups + 3xs/2xd + 1x-*O*Ac + 1x >C=CH–	24 articles/26 compounds

If the alternative argument had been used the answer would have been

5 × Me groups + 3xs/2xd + 1x-*O*Ac + 2x –CH–O–	17 articles/19 compounds

Using just ^{1}H NMR data, it was still possible to reduce the number of possible candidates to acceptable levels. Addition of the mass data (*m*/*z* 470–471) brought the selection down to one compound from one article.

6.6.2 Dereplication of a *Suberites* sp. Extract

The extract from an Antarctic *Suberites* sp. of sponge collected by SCUBA at 40 m was bioactive against the P388 cell line. The bioactive compound was isolated and partially characterized. The molecular mass was 220.04691 Da ($C_{10}H_8N_2O_4$), the λ_{max} was 348, and from the ^{1}H NMR spectrum, a 1,2,4-trisubstituted benzene system (doublet, doublet, singlet) and a trisubstituted alkene were recognizable. Notable was the lack of methyl groups in the compound. A search using *m*/*z* 220–221 Da in MarinLit returned 106 matches, but a search with $C_{10}H_8N_2O_4$ gave zero matches and established the possible novelty of the compound. To gain clues as to a possible structural type, the balance of the preliminary structural data was incorporated into a search profile:

λ_{max} 348	249 matches
λ_{max} + 0x Me groups	63 matches
λ_{max} + 0x Me + 1x 1,2,4-trisub. bz	11 matches
λ_{max} + 0x Me + 1x 1,2,4-trisub. bz + 1x > C = CH–	4 matches
λ_{max} + 0x Me + 1x 1,2,4-trisub. bz + 1x > C = CH– + *m*/*z* 220–221	0 matches

There were matches in the database right up to the point where the mass was entered. If the four matches are now examined (see Fig. 6.10), three can be quickly eliminated on the basis of disparity in mass, leaving one compound that differed by just 1 Da from a known compound: $C_{10}H_8N_2O_4$ (*m*/*z* 220) compared with $C_{10}H_9N_3O_3$ (*m*/*z* 219) for polyandrocarpamine previously isolated from the Fijian ascidian *Polyandrocarpa* sp. [30] With this structural clue, a hydantoin

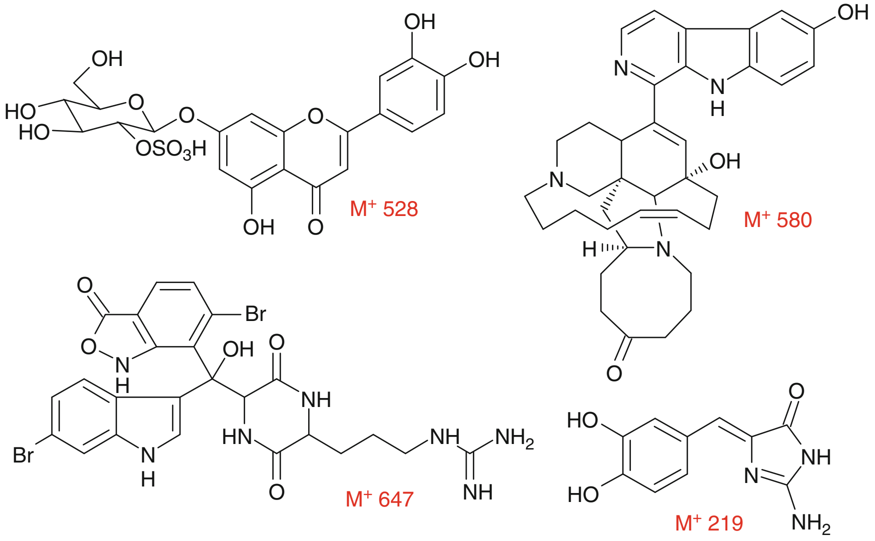
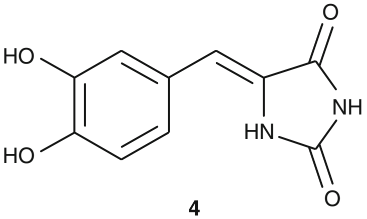


Fig. 6.10 Possible structures matching UV and NMR data from an Antarctic *Suberites* sp.

structure (4) was quickly established for the bioactive compound from the *Suberites* sp.: it was a new compound [31].



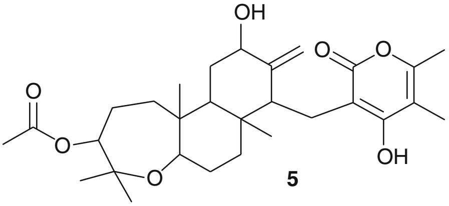
4

In this instance, just using molecular formula data was sufficient to establish that a new compound had been isolated, but interrogation of the database using other preliminary data gave essential clues as to the identity of the new compound.

6.6.3 Dereplication of an *Aspergillus* sp. Extract

Although not isolated from a marine source, the dereplication of this extract is a good example of the power of the ^{1}H-SF approach to solving problems. The ^{1}H NMR spectrum of a bioactive component isolated from the extract of an *Aspergillus* sp. isolated as an endophyte from *Garcinia scortechinii*, a Malaysian medicinal plant, contained seven singlet methyl groups (see Fig. 6.11). As the compound was isolated from a microorganism, the AntiMarin database was consulted returning just 387 possible hits out of 53,000 compounds in the database. Consideration of the chemical shift data suggested that of the seven singlet methyl groups, two were vinyl methyls and one an acetoxyl group. This refined search reduced the number of hits to five only. If the low resolution mass data (502 Da) was now added, one hit only was returned. Comparison of the ^{1}H NMR data with published data for this compound (**5**) [32] established that they were identical and completed the dereplication. The structural elucidation of three further isomers was then trivial based on the established core structure of this unusual triterpene-pyrone [33].

Alternative approaches would have used the low resolution mass data first. That would have given 91 hits reducing to just two if seven singlet methyls were included in the profile of which only one would have matched the ^{1}H NMR data obtained.



5

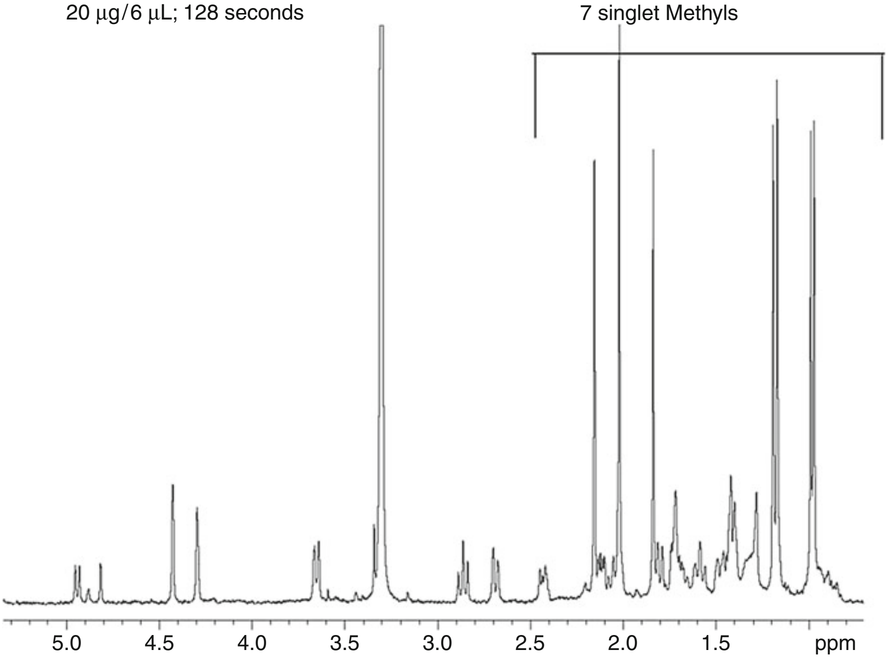


Fig. 6.11 ^{1}H NMR spectrum for a triterpene pyrone isolated from an *Aspergillus* sp. endophyte

6.7 Commentary on Approaches to Dereplication

In the section above, various approaches to achieving resolution in the dereplication process were considered. The obvious starting point is normally the molecular mass and the molecular formula, but with over 160,000 known natural products, this is not often discriminatory. Adding in taxonomic data can help narrow the dereplication to a class of compound. UV spectral data is a powerful tool in the dereplication process but is not discriminatory as it is the recognition of the chromophore, not the molecule that is occurring, and not all compounds contain chromophores. Fragmentation patterns from mass spectrometry also provide a powerful approach but require experience and skill in interpretation. The MS approach, like that of UV spectroscopy, suffers from a lack of appropriate searchable databases. The ultimate goal in dereplication is full structure determination, but to accomplish that for each and every compound in a crude extract is not a satisfactory approach and requires acquisition and interpretation of full sets of 1D and 2D NMR data in addition to relevant mass and UV data. Such a conventional approach to dereplication is shown diagrammatically in Fig. 6.12a. The alternative approach, as outlined above, is to more productively use a *minimal* set of data that helps define a structure. The recognition that a search

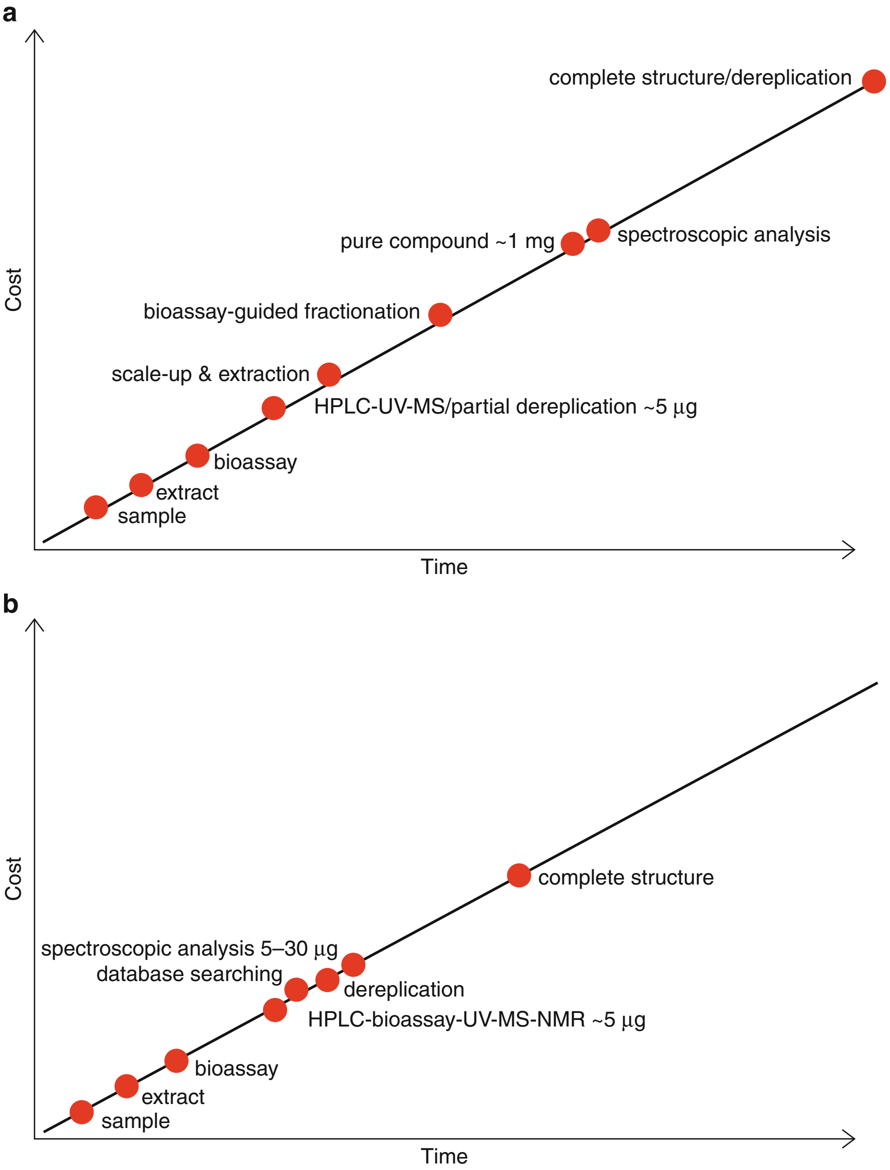
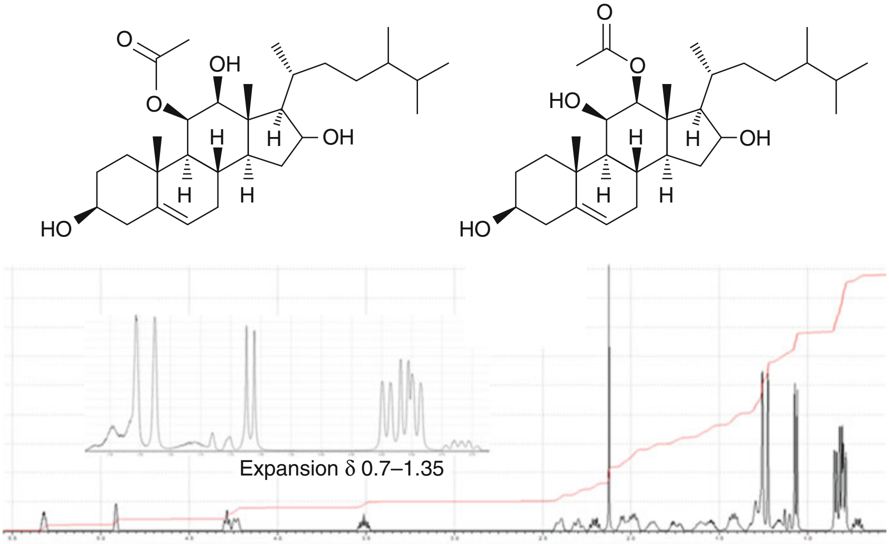


Fig. 6.12 (**a**) Conventional approach to dereplication; (**b**) Dereplication based on ^{1}H-SF approach and nanomole-scale NMR determinations

based on the numbers of functional groups easily recognizable by ^{1}H NMR spectroscopy was a powerful method for discriminating between alternative structures was the starting point for the development of the ^{1}H structural features aspect of MarinLit and subsequently AntiMarin. These are the only two databases that have such functionality. This approach allows dereplication to be accomplished and novelty established shortly after the ^{1}H NMR spectrum has been obtained and before a full interpretation of the data (see Fig. 6.12b).

This ^{1}H-SF–based approach to dereplication is well illustrated in one last example. Two isomeric compounds of molecular mass 490.3658 Da, corresponding to the molecular formula $C_{30}H_{50}O_5$ were isolated from a soft coral. Use of the various databases to look for possible structures based on this mass data has already been commented on in Table. 6.3. The ^{1}H NMR data and interpretation for one of the isomers is shown in Fig. 6.13. Without reference to mass data and simply relying on methyl group count and type, the number of possible hits in MarinLit was reduced to 43. If other information was then used, such as the four 1-H carbinol protons (δ 3.5–4.9) and the trisubstituted alkene (1-H; δ 6.35), the hits were progressively reduced to three and then two hits which corresponded to the 11-acetoxy and 12-acetoxy isomers shown in Fig. 6.13, which were first isolated from the soft coral *Capnella lacertiliensis* in 2003 [34].

If mass data had been used with the NMR interpretation, the same definitive result would have been achieved but with less NMR interpretation ($C_{30}H_{50}O_5$ gave 35 hits; $C_{30}H_{50}O_5$ + 3x Me singlets + 4x Me doublets gave 7 hits). Either approach would have lead to a full and definitive answer as to whether these were new compounds or not. The actual assignment of structure to the isomers would then be by comparison against the original data. So as with the other cases examined, dereplication has



Number of Me?	7	*Hits: 1407*	Types of Me? 3s 4 d	*Hits: 87*
Types of Me?	3 s 4 d 1 Ac	*Hits: 43*		
Other information?	4 CH-O	*Hits: 3*	Trisub alk.	*Hits: 2*

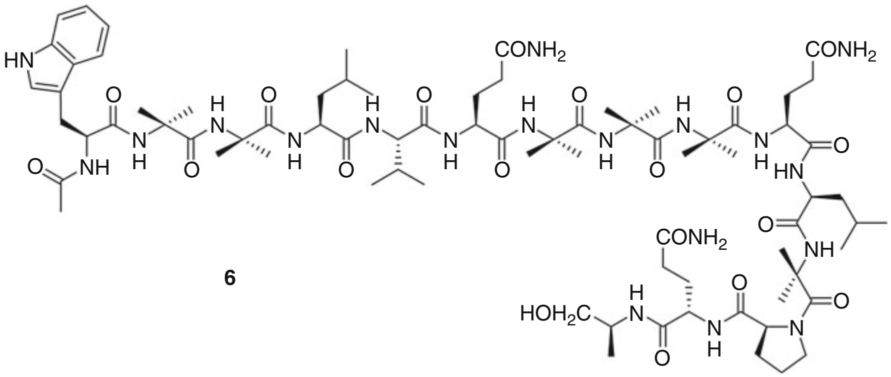
Fig. 6.13 NMR-based dereplication strategy for compounds isolated from a soft coral

been carried through quickly and efficiently and did not rely on a full structural assignment.

6.8 Dereplication at the Nanomole Level

The recent advances in NMR probe design has led to cryoprobes [35] and capillary flow probes [36] that yield 1D and 2D NMR spectra with excellent signal/noise ratios on just 2–20 μg of sample in a matter of minutes for a ^{1}H NMR spectrum to an hour or so for the likes of a COSY NMR to several hours only for an HMBC spectrum. An example of such a ^{1}H NMR spectrum is shown above in Fig. 6.11. This spectrum was obtained on 20 μg of material in 6 μL of CD_3OD in less than 2 min. This enormous advance in the relative sensitivity of the NMR experiment quickly led to numerous papers on compounds isolated at the μg level and has led to the description of a microtiter plate–based dereplication built around a Protasis CapNMR probe [37] and the MarinLit/AntiMarin databases [12, 14]. Essentially, 200–500 μg of extract are injected onto a RP-18 analytical HPLC column using an acetonitrile/water gradient (10–70% acetonitrile over 22 min) with the effluent from the column collected into 88 wells (250 μL/well) after UV and ELSD monitoring. Two daughter plates are prepared by removing 5 μL/well for biological testing and mass spectrometry. The master plate is dried and after bioassay of the daughter plate the MS and ^{1}H NMR spectra of the bioactive well(s) are obtained. These data can be immediately searched in databases and dereplication achieved. If a new compound has been detected, then further NMR data as necessary can be collected while the sample is still in the probe which meets the optimal pathway suggested in Fig. 6.12b. If a known compound had been detected, the work could be halted at that juncture.

An example of this approach would be the isolation, characterization, and structural elucidation of a new peptaibol, chrysaibol (**6**) from an extract of the fungus *Sepedonium chrysospermum* [38]. This work was carried out with an estimated 30 μg of chrysaibol isolated during HPLC analysis into the microtiter plate. This included obtaining 2D NMR data. Dereplication on the nanomole scale is possible and practicable using the database-assisted processes described in this chapter.



6.9 Relative Costs of the Databases

The cost of database searching is a real consideration. The larger databases such as Reaxys or the various version of CAS, such as SciFinder, are particularly expensive with the actual cost calculated based on factors such as the number of licenses and location. Such databases are usually paid for by central libraries at institutions rather than by individual groups. The specialized natural product databases cost considerably less and, in the main, are initially more useful for the natural product chemist. Estimates of the relative costs of the relevant databases are given in Table 6.5. Gaining access to relevant databases is not cheap, but efficient dereplication procedures can save considerable time and circumvent wasted effort leading to the more efficient throughput of samples by the researchers which is a money saver. Figure 6.12a and b attempt to depict this aspect of efficient dereplication in terms of a time/cost exercise. In Corley's 1994 paper on the strategies for database dereplication, he estimated that "in our laboratory that for each natural product dereplicated, at an average cost of $300 of online time (using STN databases), a savings of $50,000 is incurred in isolation and identification time." [39] If that was a true cost/benefit analysis in 1994, imagine the benefits in 2012 and into the future? Databases play an essential role in the dereplication of natural product extracts (Table 6.5).

6.10 Study Questions

1. What are the advantages and disadvantages of using a specialist database to extract specific data in a narrow field such as marine natural products as opposed to using a generalist, all encompassing database such as CAS Registry?

Table 6.5 Relative costings of the databases useful in natural products research

		Number of compounds	
Database	Cost	Total	Natural products
SciFinder	>US$50,000	6.3×10^7	~260,000
CSLS	free	4.6×10^7	extracts
ChemSpider	free	2.6×10^7	?
PubChem	free	3.7×10^7	?
Reaxys	>US$40,000	$>10^7$	170,000
NaprAlert	US$0.5/citation		>150,261
Dictionary of NaturalProducts	US$6,600		159,670
AntiMarin[a]	$0		53,000
AntiBase	US$3,500		36,000
Dictionary of Marine Natural Products	US$625		22,691
MarinLit	US$1,850		22,000

[a] To obtain a copy of AntiMarin it is first necessary to be a current subscriber to the AntiBase and MarinLit databases

2. In the dereplication of a natural product extract, suggest a minimal data set that would establish the uniqueness of each compound isolated.
3. What are the pitfalls that may be encountered when using taxonomic data in a dereplication exercise?
4. Outline the problems that would be encountered if molecular mass or molecular formulae only were used in a dereplication exercise?
5. Suggest reasons why a database that includes NMR characteristics for each compound is more likely to be discriminating than any based on other spectral and physical properties such as mass, molecular formulae, UV, MS fragmentation, or IR.

References

1. Buckingham J (ed) (2011) Dictionary of natural products on DVD. Chapman & Hall/CRC, Boca Raton
2. CAS, www.cas.org. Accessed 18 Nov 2011
3. Reaxys, https://www.reaxys.com/info/. Accessed 18 Nov 2011
4. CAS Registry, www.cas.org/expertise/cascontent/registry/index.html. Accessed 18 Nov 2011
5. STN, http://www.stn-international.de/index.php?id=123. Accessed 18 Nov 2011
6. SciFinder, http://www.cas.org/products/scifindr/. Accessed 18 Nov 2011
7. ChemNetBase, http://www.chemnetbase.com/. Accessed 18 Nov 2011
8. Blunt JW, Munro MHG (eds) (2008) Dictionary of marine natural products. Chapman & Hall/CRC, Boca Raton
9. Jaspars M (2008) Underwater chemistry. Chem World 5(Jan). http://www.rsc.org/chemistryworld/Issues/2008/January/Reviews.asp. Accessed 18 Nov 2011
10. AntiBase, http://www.wiley-vch.de/publish/dt/books/forthcomingTitles/LS00/3-527-32827-0/?sID=fdd57124490d0f87cfd550f72e2684d5. Accessed 18 Nov 2011
11. SpecInfo, http://cds.dl.ac.uk/cds/datasets/spec/specinfo/specinfo.html. Accessed 18 Nov 2011
12. MarinLit, http://www.chem.canterbury.ac.nz/marinlit/marinlit.shtml. Accessed 18 Nov 2011
13. ACD/Labs, http://www.acdlabs.com. Accessed 18 Nov 2011
14. AntiMarin: a combination database formed from AntiBase and MarinLit. http://www.wiley-vch.de/publish/dt/books/forthcomingTitles/LS00/3-527-32827-0/?sID=fdd57124490d0f87cfd550f72e2684d5 and http://www.chem.canterbury.ac.nz/marinlit/marinlit.shtml. Accessed 18 Nov 2011
15. http://depth-first.com/articles/2011/10/12/sixty-four-free-chemistry-databases/. Accessed 18 Nov 2011
16. Williams AJ (2008) Public chemical compound databases. Current Opin Drug Discov Develop 11:393–404
17. PubChem, http://pubchem.ncbi.nlm.nih.gov/. Accessed 18 Nov 2011
18. CSLS, http://cactus.nci.nih.gov/lookup/. Accessed 18 Nov 2011
19. ChemSpider, http://www.chemspider.com/. Accessed 18 Nov 2011
20. NaprAlert, http://www.napralert.org/. Accessed 18 Nov 2011
21. Perry NB, Ettouati L, Litaudon M et al (1994) Alkaloids from the Antarctic sponge *Kirkpatrickia varialosa* Part 1: Variolin B, a new antitumor and antiviral compound. Tetrahedron 50:3987–3892
22. Nielsen KF, Smedsgaard J (2003) Fungal metabolite screening: database of 474 mycotoxins and fungal metabolites for dereplication by standardized liquid chromatography-UV-mass spectrometry methodology. J Chrom A 1002:111–136

23. Dumdei EJ, Blunt JW, Munro MHG et al (1997) Isolation of the calyculins, calyculinamides and swinholide H from the New Zealand deep water sponge *Lamellomorpha strongylata*. J Org Chem 62:2636–2639
24. The Marine Group, University of Canterbury's UV data was acquired on a Dionex HPLC using Chromeleon software. The library of known compounds is available on application to murray.munro@canterbury.ac.nz or john.blunt@canterbury.ac.nz
25. NMR Shift DB, http://nmrshiftdb.nmr.uni-koeln.de/. Accessed 18 Nov 2011
26. Naproc-13, http://c13.usal.es/c13/usuario/views/inicio.jsp?lang=en&country=EN. Accessed 18 Nov 2011
27. SuperNatural, http://bioinformatics.charite.de/supernatural/. Accessed 18 Nov 2011
28. SDBS (Spectral Database for Organic Compounds), http://riodb01.ibase.aist.go.jp/sdbs/cgi-bin/cre_index.cgi?lang=eng. Accessed 18 Nov 2011
29. Wellington KD, Cambie RC, Rutledge PS et al (2000) Chemistry of sponges 19. Novel bioactive metabolites from Hamigera taragensis. J Nat Prod 63:79–85
30. Davis RA, Aalbersberg W, Meo S et al (2002) The isolation and synthesis of polyandrocarpamines A and B. Two new 2-aminoimidazolone compounds from the Fijian ascidian, Polyandrocarpa sp. Tetrahedron 58:3263–3269
31. MacLean WJ (2005) Unpublished results. PhD thesis. University of Canterbury, Christchurch
32. Suzuki K, Kuwahara A, Nishikiori T et al (1997) NF00659A1, A2, A3, B1 and B2, novel antitumor antibiotics produced by *Aspergillus* sp. NF 00659. II. Structural elucidation. J Antibiot 50:318–324
33. Sun Lin (2009) Unpublished work. PhD thesis. University of Canterbury, Christchurch
34. Wright AD, Goclik E, Konig GM (2003) Oxygenated analogues of gorgosterol and ergosterol from the soft coral *Capnella lacertiliensis*. J Nat Prod 66:157–160
35. See for example http://www.bruker-biospin.com/cryoprobes.html, or http://www.varianinc.com/cgi-bin/nav?products/nmr/probes/liquids/cold_probes/index&cid=LNMJOLHKFI. Accessed 18 Nov 2011
36. http://www.protasis.com/OMNMR/index.htm. Accessed 18 Nov 2011
37. Lang G, Mayhudin NA, Mitova MI et al (2008) Evolving trends in the dereplication of natural product extracts: new methodology for rapid, small-scale investigation of natural product extracts. J Nat Prod 71:1595–1599
38. Mitova MI, Murphy AC, Lang G et al (2008) Evolving trends in the dereplication of natural product extracts. 2. The isolation of chrysaibol, an antibiotic peptaibol from a New Zealand sample of the mycoparasitic fungus *Sepedonium chrysospermum*. J Nat Prod 71:1600–1603
39. Coreley DG, Durley RC (1994) Strategies for database dereplication of natural products. J Nat Prod 57:1484–1490
40. GVK Biosciences Natural Product DB, http://www.gvkbio.com/db_products.html. Accessed 18 Nov 2011
41. Lei J, Zhou J (2002) A marine natural product database. J Chem Info Comput Sci 42:742–748
42. Bitzer J, Kopcke B, Stadler M et al (2007) Accelerated dereplication of natural products, supported by reference libraries. Chimia 61:332–338
43. Moss S, Bovermann G, Denay R et al (2007) Efficient structure elucidation of natural products in the pharmaceutical industry. Chimia 61:346–349
44. The National Centre for Plant and Microbial Metabolomics, http://www.metabolomics.bbsrc.ac.uk/currentactivities.htm. Accessed 18 Nov 2011
45. CH-NMR-NP, http://www.las.jp/english/software/chnmrnp.html. Accessed 18 Nov 2011