A Rapid and Facile Method for the Dereplication of Purified Natural Products

John Bradshaw,[†,‡] Darko Butina,[†,§] Adrian J. Dunn,[†] Richard H. Green,[†] Michaela Hajek,[⊥,∇] Matthew M. Jones,[†] John C. Lindon,[⊥] and Philip J. Sidebottom*,[†]

GlaxoSmithKline Medicines Research Centre, Gunnels Wood Road, Stevenage, Hertfordshire, SG1 2NY, U.K., and Biological Chemistry, Biomedical Sciences Division, Imperial College of Sciences, Technology and Medicine, Sir Alexander Fleming Building, South Kensington, London, SW7 2AZ, U.K.

Received June 15, 2001

A new approach to the use of commercial databases for the dereplication of purified natural products has been developed. This is based on searching a text file that links each structure with its molecular weight and an exact count of the number of methyl, methylene, and methine groups it contains. Analysis of such a text file, constructed from a database containing more than 126 000 natural product structures, revealed that these data, readily measured using MS and NMR spectroscopy, are highly discriminating. The identification of an alkaloid and a sesquiterpene using this new approach is described.

The rapid identification of already known natural products, a process known as dereplication, is strategically important for scientists involved in screening for novel bioactive compounds from natural sources. Efficient dereplication is essential if expensive isolation and structure elucidation resources are not to be squandered since a successful discovery program depends on these resources being focused primarily on the key samples likely to generate new natural product leads.

The most common procedures used to identify compounds prior to purification are based on LC-UV, LC-MS, or a combination of the two.[1,2] For these procedures to succeed authentic samples are required. Initially, these are used to create a suitable database of LC retention times and spectra measured under standard conditions. More importantly they enable bioactivity data to be obtained, thus allowing the identification of which compounds, if present, will account for the activity of the crude extract.

Currently, there are more than 150 000 known natural products, and new structures are being published at a rate of about 10 000 per annum.[3] The percentage of the known natural products available to any organization through compound libraries is thus likely to be small and diminishing. It follows that procedures such as those outlined above will provide only a partial solution to the dereplication problem. In the past decade, commercial databases containing information on natural products have proved to be an increasingly important way of filling this gap, particularly for purified compounds.[4]

When a compound of interest has been located in the crude extract, following bioactivity-guided fractionation for example, it has then to be identified. Data from UV and MS alone will rarely provide sufficient information to distinguish between the known isomers. Even when this is achieved, possible novel isomers have also to be considered. Recently the use of LC-NMR at this stage has been advocated.[5] While this will help in favorable cases, the absence of data from certain regions of the spectrum due to the need to suppress the large solvent peaks, and the lack of literature NMR data in the mixed solvent systems used, means that it seems unlikely to provide a general solution.

Our use of commercial databases has focused on simplifying and speeding up the structural elucidation of what turn out to be known compounds. Information from fully purified samples has the advantage that the final identification can be achieved by a comparison with literature data. To provide a small number of candidates for this comparison, substructure searching using small structural fragments (e.g., methoxy or ethyl) often in conjunction with molecular weight information has been employed. The small fragments have been identified from 1D ^{1}H and 2D, usually ^{1}H−^{13}C HMQC, NMR spectra. Experience has shown that while this is extremely useful, there are a number of problems and inefficiencies in this approach to database searching. First, as noted by Corley and Durley,[4] a number of different, quite complex software packages need to be mastered if all the available databases are to be utilized. Second, in some databases, substructure searches using a number of small fragments are prohibitively slow probably because the routines were not developed with this task in mind.[6] Finally, a substructure search for a fragment returns as hits all compounds with at least one such fragment. To obtain only those compounds with exactly one such fragment, a second search and subsequent subtraction of hit lists usually has to be carried out. These difficulties led us to look for a new procedure using data that can be both calculated directly from structures and rapidly measured for new samples. In this paper we report the results of these endeavors.

Results and Discussion

It was reasoned that a file combining structures together with an exact numerical count for each of the readily observable small fragments they contained would provide a good way forward. Such a file could be created using software from Daylight[7] and should be fast and easy to use, as only a simple text search would have to be employed. While this software handles chemical structures in SMILES format,[7] conversion routines from other common structure formats to SMILES are available. Thus, in principle, a single data file incorporating natural product structures from a variety of sources and including “in-house” data could be constructed. Determining the number and nature

* To whom correspondence should be addressed. Tel: +44 (0) 1438-763319. Fax: +44 (0) 1438-763352. E-mail: pjs3111@gsk.com.

[†] GlaxoSmithKline Medicines Research Centre.

[‡] Current address: Daylight Chemical Information Systems Inc, Sheraton House, Castle Park, Cambridge, CB3 0AX, U.K.

[§] Current address: Camitro (UK) Ltd, 127 Science Park, Milton Road, Cambridge, CB4 0GD, U.K.

[⊥] Imperial College of Sciences, Technology and Medicine.

[∇] Current address: GlaxoSmithKline Medicines Research Centre.

10.1021/np010284g CCC: $20.00
Published on Web 11/28/2001

Table 1. Analysis of the Number of Occurrences, within the Test Database of Natural Products, of the 7188 Different Combinations of the Number of CH_3, CH_2, and CH Groups

occurrences per combination	no. of combinations	no. of structures
1−10	5391	14 309
11−20	562	8269
21−30	276	7017
31−40	171	6004
41−50	123	5494
51−60	118	6449
61−70	78	5113
71−80	61	4589
81−90	48	4102
91−100	43	4089
101−461	317	60 678

of the fragments that would provide the necessary discrimination was the first task.

In the absence of a commercially available Daylight database of natural product structures, one constructed “in-house” for other purposes was used to test this concept. It contained more than 126 000 unique SMILES structures derived from the commercial databases, Chapman & Hall’s *Dictionary of Natural Products*, and Beilstein. It is important to note that as stereochemistry was not given in the source files used to create the Daylight database, the SMILES we have used does not contain this information. As a result, some of the entries in the database contain data on more than one stereoisomer.

Initially, the three simplest fragments (CH_3, CH_2, and CH) were counted. Analysis of the resulting file showed that there were 7188 different combinations of these fragments. Of these, 5391 combinations were exclusive to 10 structures or less. Even the most common combination, two methyls, zero methylenes, and six methines (2−0−6), only occurred 461 times. More detailed analysis (Table 1) showed that this initially encouraging result was not a complete solution, as nearly half of the database was represented by just 317 combinations.

Rather than introduce more fragments, the next step was to incorporate the molecular weight into the search strategy. To do this, it was necessary to calculate the molecular weights from the structures, as the MW field could not be exported from Beilstein. To fit in with the low-resolution MS data routinely available, it was decided to calculate the integer monoisotopic masses using the mass of the most abundant isotope of each element.

While the molecular weights of compounds in the database extended beyond 3500, only about 2.5% of the entries were bigger than 1000. More than 80% of the compounds had molecular weights in the range 200−700. However, even the most frequently occurring mass accounted for less than 1% of the database. This preliminary analysis suggested that molecular weight information would indeed complement the fragment count approach. Clearly for larger molecules (MW > 1000) counting the number of CH_3, CH_2, and CH groups may be difficult. In these cases molecular weight alone should be enough to achieve the objective of producing only a few structures for consideration. To confirm that molecular weight was also useful for smaller molecules, we analyzed their distribution within the sets of compounds having the 10 most common fragment combinations. This showed, for example, that the 461 structures with a 2−0−6 combination had 156 different molecular weights extending up to above 900. All but four of these occurred less than 11 times. Broadly similar results were obtained for the other combinations (Table 2).

This confirmed that we now had a procedure that would quickly reduce 126 000 possibilities, most times to a handful, and probably every time to less than 70. For most compounds the molecular weight can be acquired readily by mass spectrometry using the widely available electrospray or APCI ionization methods. An accurate count of the number of CH’s, CH_2’s, and CH_3’s can be obtained most easily from $^1H-^{13}C$ correlation data, measured by inverse-detection NMR methods such as HMQC or HSQC, together with a careful analysis of the ^1H NMR spectrum. These NMR data can be collected within a couple of hours on a few milligrams of compound using a standard modern NMR spectrometer and much faster if the latest technology is available.[8] Alternatively this information can be derived routinely from DEPT ^{13}C NMR spectra, although this will generally require either more compound or more spectrometer time.

To see if a further simple improvement was possible, we examined the structures of the compounds making up the four groups with more than 40 members (Table 3). This revealed that each group was a series of closely related isomers of which nearly all were members of the flavonoid class. Clearly, employing an additional count of the quaternary carbons, as used in a recent program aimed at the automatic identification of terpenoid skeletons,[9] would be futile in distinguishing within these groups of isomers. Similarly, HRMS would provide no extra discrimination in these cases. While both these additions would provide extra discrimination in other cases, they would also be much more difficult to measure than the data used in our proposed approach. If further discrimination of these flavonoid isomers were required, then fragments related to aromatic substitution patterns would appear to provide the best way forward.

The following worked examples serve to demonstrate how the approach was applied in practice. Through a research collaboration, we received a sample of an uniden-

Table 2. Analysis of the Distribution of Molecular Weights within Each of the Compound Sets Having the 10 Most Common CH_3-CH_2-CH Combinations

			no. of molecular weights which occur					
no. of CH_3-CH_2-CH	no. of occurrences	no. of different MWs	1−5 times	6−10 times	11−20 times	21−30 times	31−40 times	41−70 times
2−0−6	461	156	140	12	3	0	0	1
2−0−4	460	138	121	11	3	1	2	0
1−0−5	445	136	114	17	3	1	1	0
0−0−6	432	181	169	7	4	0	0	1
1−1−5	429	161	143	16	2	0	0	0
0−0−5	407	150	138	8	2	1	0	1
1−0−4	400	120	106	6	6	2	0	0
0−0−4	399	156	143	9	2	1	1	0
2−0−5	399	123	107	11	4	0	0	1
3−1−5	397	137	131	6	0	0	0	0

Table 3. Groups of Compounds with More than 40 Members

no. of CH_3-CH_2-CH	molecular weight	molecular formula	no. of compounds
2−0−6	314	$C_{17}H_{14}O_6$	65
0−0−6	286	$C_{15}H_{10}O_6$	54
0−0−5	302	$C_{15}H_{10}O_7$	56
2−0−5	330	$C_{17}H_{14}O_7$	68

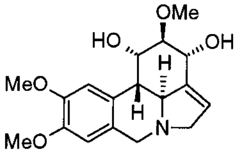
Table 4. NMR Data (1D ^1H and HMQC) in $CDCl_3$ for the Unknown Alkaloid with a Molecular Weight of 333

δ ^{13}C	δ ^1H
41.2	2.74 (1H, dd, J = 2.0, 11.0 Hz)
54.0	4.10 (1H, d, J = 13.0 Hz)
	3.59 (1H, d, J = 13.0 Hz)
55.6	3.84 (3H, s)
55.6	3.89 (3H, s)
57.6	3.46 (3H, s)
62.0	3.87 (1H, m)
62.1	4.07(1H, dt, J = 14.5, 2.0 Hz)
	3.60 (1H, ddd, J = 2.0, 5.5, 14.5 Hz)
67.9	4.70 (1H, m)
68.9	4.69 (1H, d, J = 3.0 Hz)
80.0	3.81 (1H, t, J = 3.0 Hz)
107.3	6.90 (1H, s)
110.2	6.70 (1H, s)
120.9	5.61 (1H, m)

tified alkaloid having a mass of 333. After recording 1D ^1H and HMQC NMR spectra (Table 4) we were able to count the number of CH_3's, CH_2's, and CH's (3−2−8).

Although not relevant in this case, it is important to consider symmetry. An isopropyl group, for example, contains two methyl groups that may give only one double intensity signal. Careful attention to the integrals from the proton spectrum should enable such cases to be counted correctly. When the molecular weight is more than double that which the combination requires,[10] the possibility of a fully symmetrical dimer or trimer cannot be excluded. The speed of the search, which takes seconds rather than minutes, allows a pragmatic solution to this problem; the search is rerun using double or triple the counts.

Running the search (3−2−8, MW = 333) gave a single hit, narcissidine (**1**). The next step was to eliminate any of the suggested structures that were not consistent with the NMR data. With a little practice, simple checks, based on comparing the expected and observed chemical shifts and multiplicities, enable this to be carried out very quickly. Narcissidine passed all such checks in relation to the data in Table 4.



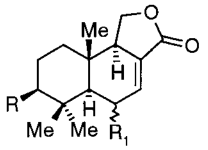
1

The final step was to locate the published data for the structures that remained and to compare them with that for the unknown. This was achieved by searching the original database(s) using the unique identifiers stored in the Daylight database. Thus, the BRN number was used to search Beilstein and the UKEY to search the *Dictionary of Natural Products*. These simple text searches were both simple to carry out and very fast to run. They gave ready access to the literature references for a range of published data. The published NMR data for narcissidine (**1**)[11] was in excellent agreement with that given in Table 4.

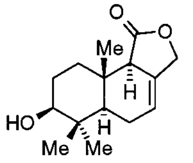
Table 5. NMR Data (1D ^1H and HMQC) in $CDCl_3$ for the Unknown Compound with a Molecular Weight of 250

δ ^{13}C	δ ^1H
13.4	0.80 (3H, s)
14.9	0.91 (3H, s)
24.8	2.21 (1H, dddd, J = 20.0, 11.5, 5.0, 3.5 Hz)
	2.44 (1H, ddt, J = 20.0, 5.0, 3.5 Hz)
26.9	1.64 (1H, m)
	1.69 (1H, m)
27.8	1.04 (3H, s)
37.3	1.34 (1H, td, J = 13.5, 4.0 Hz)
	1.64 (1H, m)
49.1	1.37 (1H, dd, J = 11.5, 5.0 Hz)
50.6	2.78 (1H, m)
67.0	4.04 (1H, t, J = 9.0 Hz)
	4.38 (1H, t, J = 9.0 Hz)
78.4	3.30 (1H, dd, J = 4.5, 11.0 Hz)
136.1	6.89 (1H, q, J = 3.5 Hz)

A second unrelated compound, with a mass of 250, gave the NMR data listed in Table 5. This was a more challenging example in that the initial search (3−4−4, MW = 250) yielded 40 hits, all having the molecular formula $C_{15}H_{20}O_3$. However, of these candidate structures only 15 contained the correct number of proton-bearing sp^2 carbons. Of these 15 only nine contained a CH_2-O group, and only three of these nine would give rise to three singlets resonating at below 1.1δ in the ^1H NMR spectrum. As two stereoisomers of one of these structures were known, four candidate structures (**2**−**5**) remained at this point of the analysis.



2 R = OH, R_1 = H
3 R = H, R_1 = α OH
4 R = H, R_1 = β OH



5

Comparison with the literature NMR data for these enabled the unknown to be unequivocally identified as 3β-hydroxycinnamolide (**2**).[12]

If at any stage of the process no structures remain, then we conclude that the unknown compound may be novel[13] and hence worthy of further spectroscopic and, if necessary, isolation work.

It is important to realize that the benefit of using this approach to the dereplication of purified natural products extends beyond improving the efficiency of the structure elucidation process. With modern instrumentation the data that are needed can be rapidly measured on sub-milligram amounts of compound.[8] This should enable the initial isolation of a bioactive compound to be carried out using the original screening sample. A larger scale re-fermentation or re-extraction should only be needed if the compound is present at very low concentration, and hence can be assumed to be very active, or when the compound appears novel and requires the use of less sensitive NMR experiments for its structure elucidation.

Experimental Section

General Experimental Procedures. The procedures given below can be carried out on any computer with a UNIX-based operating system and the necessary Daylight software. The times quoted are those obtained using a Silicon Graphics O2 fitted with a mips R5000 300 MHz IP32 processor and 384Mb of memory.

The Daylight database used was built using two SDF files. The first, of the *Dictionary of Natural Products* (version 6.2; 1998), was purchased directly from Chapman and Hall.[14] The second, containing compounds flagged as "Isolation from Natural Product", was downloaded with permission, via CrossFire, from the Beilstein database (version 9801PR; 1998).[15]

A text file containing a list of SMILES, one per line, was generated from the Daylight database. The fragments to be counted were defined using SMARTS, an extension of SMILES used when defining substructures. The definitions used for methyl, methylene, and methine groups were [CH3], [CH2], and [#6;H1], respectively.

A script written using the Daylight tool kit was used, in conjunction with a file containing the fragment definitions, to count the list of SMILES. For each line it calculated and appended the number of occurrences of each of the defined fragments within the SMILES. This step took less than 10 min. The output file was further processed to add the molecular weights. These were calculated directly from the SMILES.

The resulting text file was searched using the "awk" function of UNIX. This has sufficient flexibility to allow the simultaneous search of multiple fields each for either a specific value or a range of values. The output is a file of the SMILES from the lines that match the search criteria. Search times were typically measured in seconds.

The matching structures could be viewed using the Daylight "depict" routine. The associated data from the original Daylight database, including the UKEY and BRN numbers, could be accessed using the Daylight "thorlookup" routine. In practice, the intricacies of the search and display process can be hidden behind an easy to use Web-based front end.

NMR spectra were recorded on a Bruker AMX 500 spectrometer equipped with a Nalorac 3 mm ^{1}H/BB probe fitted with Z gradient coils. The HMQC data were acquired with 256 × 1K data points using a standard pulse sequence[16] in 36 min.

Acknowledgment. We thank Dr. Fiona Macdonald and Dr. S. P. Walford from Chapman & Hall for providing a modified dnp.sf1 file that enabled us to search the UKEY field in the CD-ROM version of the *Dictionary of Natural Products*. The alkaloid sample was provided by Prof. Monique Simmonds from the Royal Botanical Gardens at Kew. A postdoctoral research fellowship from GlaxoSmithKline to M.H. is acknowledged.

Supporting Information Available: The structures of the 65 compounds in the 2−0−6 group with MW = 314 (see Table 3). The structures of the 40 compounds in the 3−4−4 group with MW = 250. An example of part of the text file containing SMILES, MW, and counts of CH_3's, CH_2's, and CH's. This material is available free of charge via the Internet at http://pubs.acs.org.

References and Notes

(1) Hook, D. J.; Pack, E. J.; Yacobucci, J. J.; Guss, J. *J. Biomol. Screening* **1997**, *2*, 145−152.
(2) Cordell, G. A.; Shin, Y. G. *Pure Appl. Chem.* **1999**, *71*, 1089−1094.
(3) The number of entries in the Chapman & Hall *Dictionary of Natural Products* rose from 108 000 in 1995 to 148 000 in 1999.
(4) Corley, D. G.; Durley, R. C. *J. Nat. Prod.* **1994**, *57*, 1484−1490.
(5) Bobzin, S. C.; Yang, S.; Kasten, T. P. *J. Ind. Microbiol. Biotech.* **2000**, *25*, 342−345.
(6) Ozawa, K.; Yasuda, T.; Fujita, S. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 688−695, and references therein.
(7) For details of software available from Daylight, Inc. and the SMILES format see their web site: www.daylight.com.
(8) Russell, D. J.; Hadden, C. E.; Martin, G. E.; Gibson, A. A.; Zens, A. P.; Carolan, J. L. *J. Nat. Prod.* **2000**, *63*, 1047−1049, and references therein.
(9) Ferreira, M. J. P.; Brant, A. J. C.; Rodrigues, G. V.; Emerenciano, V. P. *Anal. Chim. Acta* **2001**, *429*, 151−170.
(10) For a methyl−methylene−methine combination of $x-y-z$ a MW ≥ $15x + 14y + 13z$ is required.
(11) Kihara, M.; Ozaki, T.; Kobayashi, S.; Shingu, T. *Chem. Pharm. Bull.* **1995**, *43*, 318−320.
(12) Ayer, W. A.; Trifonov, L. S. *J. Nat. Prod.* **1992**, *55*, 1454−1461.
(13) Clearly, commercial databases will not include recently published structures. This lag may in some cases be more than a year.
(14) Chapman & Hall is now part of the CRC Press Group. For details see their web site: www.crcpress.com.
(15) For details of the Beilstein database and CrossFire software see their web site: www.beilstein.com.
(16) Bax, A.; Subramanian, S. *J. Magn. Reson.* **1986**, *67*, 565−569.

NP010284G