Techniques for Generating Descriptive Fingerprints in Combinatorial Libraries[§]

Geoff M. Downs[†] and John M. Barnard*,[‡]

Barnard Chemical Information Ltd., 46 Uppergate Road, Stannington, Sheffield S6 6BX, UK

Received June 27, 1996[⊗]

One of the problems encountered in handling computer representations of combinatorial libraries (especially "virtual libraries") is the extremely large number of compounds which may be covered by even quite simple libraries. This paper discusses work on the generation of structure fingerprints for the compounds in a library, which avoids the need to enumerate the compounds themselves by utilizing techniques originally developed for handling Markush structures in chemical patents.

INTRODUCTION

A single combinatorial library (CL) of compounds may contain a very large number of individual compounds (10^6, 10^9, or more). This can cause problems for computer systems which depend on enumeration of these compounds in order to be able to perform any sort of analysis of their diversity. Problems such as clustering the compounds in a large CL, which may have time and/or space requirements proportional to the square of the number of the number of compounds, are thus too slow to be practicable. Though it is now generally considered that only relatively modest CLs (around 10^3-10^4 compounds) will actually be physically synthesized, computer systems will still be required to handle "virtual" CLs several orders of magnitude larger than this. Indeed, performing diversity analyses on such virtual CLs may be an essential step in deciding the optimum subset CL which is actually to be synthesized. Recent work[1] has focused on considering the diversity of each building block pool in isolation, and though the nature of combinatorial synthesis requires that subsets of each building block pool (rather than random combinations of members of different pools) must be chosen in order to build a subset CL, this approach presupposes that maximizing the diversity in each pool subset will maximize the diversity of the compounds in the subset CL.

A similar problem of handling extremely large sets of compounds occurs in systems for the storage and retrieval of Markush structures in chemical patents. By inclusion of expressions such as "R1 is an optionally-substituted heterocycle", Markush structures in patents frequently cover open-ended (i.e., infinite) sets of compunds; this clearly rules out enumeration as a technique for dealing with them, and other means have had to be employed. A long-running research project at Sheffield University[2] developed a number ot techniques for representing and searching databases of Markush structures, without enumeration of their coverage, and several operational systems have been developed commercially, for use with patent databases.[3]

CLs can clearly be considered as a type of Markush structure, though they are generally much simpler than those found in patents. Simple lists of alternative specific substituents on a central scaffold predominate, and "open-ended" sets are entirely absent, though it might be argued that they are potentially useful in description of extremely large virtual CLs.

Some of the techniques used for searching databases of Markush structures from patents are also being employed in the search systems currently under development by major software vendors for CL information systems. At the simplest level, this involves storing and matching the common parts of the compounds covered by a CL only once, in order to avoid full enumeration. Such techniques are not especially new, and the earliest such system was designed as long ago as 1958.[4] The work described in this paper examines the application of Markush structure-handling techniques to diversity analysis within CLs, in order to avoid exhaustive enumeration of the compounds covered.

INITIAL AIMS OF WORK

Following conference presentations during 1995, on the relationship between the Markush structures found in patents and those used to describe CLs,[5] work was commenced with the following initial aims:

1. to design an internal data structure for Markush structures suitable for representing CLs,
2. to generate structure "fingerprints" (indicating the presence or absence of certain substructural fragments in each compound) for all the compounds in a CL directly from this,
3. to compare these fingerprints with those obtained by first enumerating the compounds and then generating fingerprints for each individually.

Though this process still involves enumeration of structure fingerprints (which might then be used for clustering, dissimilarity selection etc.) the initial purpose was to demonstrate the applicability of the approach, which in any case may offer significant savings in processing time; extension of the work to avoid even this enumeration is discussed at the end of this paper.

Figure 1 shows the basic processing flow involved. Two alternative routes were used to generate structure fingerprints from initial input of a CL representation: on the left-hand side, a file containing the enumerated connection tables for all the compounds in the CL is generated, which can then be used as input to standard fingerprint generation software. On the right-hand side new software is used to read the CL

[§] Presented at the 4th International Conference on Chemical Structures, Noordwijkerhout, The Netherlands, June 2−6, 1996.

[†] E-mail address: downs@bci2.demon.co.uk.

[‡] E-mail address: barnard@bci1.demon.co.uk.

[⊗] Abstract published in *Advance ACS Abstracts,* December 15, 1996.

S0095-2338(96)00091-1 CCC: $14.00

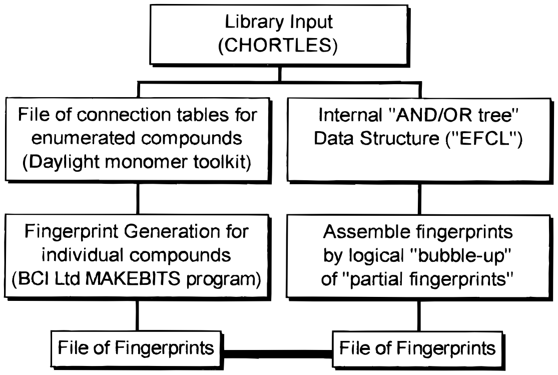
Library Input (CHORTLES)
File of connection tables for enumerated compounds (Daylight monomer toolkit)
Internal "AND/OR tree" Data Structure ("EFCL")
Fingerprint Generation for individual compounds (BCI Ltd MAKEBITS program)
Assemble fingerprints by logical "bubble-up" of "partial fingerprints"
File of Fingerprints
File of Fingerprints

Figure 1. Overall processing flow for two methods of fingerprint generation for combinatorial libraries. Explanation in text.

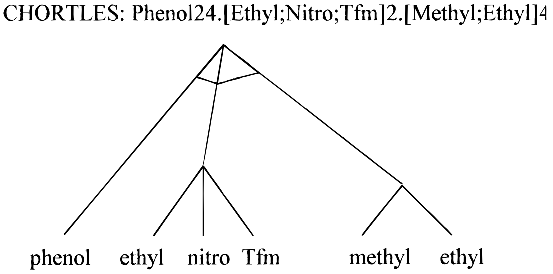

Figure 2. Schematic representation of the structure of the EFCL represention for a simple combinatorial library, also shown as a CHORTLES notation. Each "leaf node" in the tree is a partial structure, representing one building block (monomer unit). The lines connected by small arcs indicate AND relationships between the partial structures at the ends of the lines, whilst those without arcs represent OR relationships.

into the specially-designed internal representation and then to generate the fingerprints directly from it. The fingerprint files on each side should, of course, be identical.

INTERNAL REPRESENTATION OF CLS

In the Sheffield research work, a specially-designed internal representation, called the Extended Connection Table Representation (ECTR)[6] was used for Markush structures, and two important algorithms were developed to process it. The ECTR represents a Markush structure as a logical "AND/OR" tree in which leaf nodes represent individual alternatives for each variable group ("partial structures"), and internal nodes show the logical relationships between them. The TreeTrace algorithm traverses the ECTR accessing each partial structure in turn, and the Bubble-Up algorithm accumulates information from each partial structure, keeping the correct logical relationships, and "bubbles it up" to the top (root) of the tree.[7]

For the present work, a conceptually similar data structure was designed, called "ECTR For Combinatorial Libraries" (EFCL). This is implemented using dynamic arrays in the C language and is intended to be created and held in random-access memory only for as long as is required for processing; no arbitrary limits are placed on the number of partial structures, the number of connections between any pair of partial structures, or the depth of the tree. Figure 2 illustrates the logical structure of the EFCL with a simple example.

FINGERPRINT GENERATION

For reasons of availability and convenience, significant use was made of the Monomer Toolkit supplied by Daylight Chemical Information Inc.[8] CLs were represented as CHORTLES notations,[9] with associated building block definitions. Daylight Monomer Toolkit routines were used to enumerate the individual compound connection tables which were then used as input to Barnard Chemical Information Ltd.'s standard fingerprint generation program, MAKEBITS.[10] The Daylight fingerprint generation routines were not used, as these could not have been used for fingerprint generation directly from the EFCL.

Table 1. Processing Timings (Elapsed Seconds) for Generation of Files of Structure Fingerprints from the Three Sample CLs, Using the Enumeration and Markush Routes[a]

library	no. of compds	enumeration method	Markush method
PAM95	600	49.6s	1.23s
PEPTIDE95	1280	98.0s	5.77s
PEPTOID95	1280	116.0s	5.67s

[a] Programs were run on a Silicon Graphics Indigo 2 computer.

Special routines (also using the Daylight Monomer Toolkit) were written in order to generate the EFCL from CHORTLES input, though there is no reason why it should not also be generated from other input formats such as Tripos' Combinatorial Sybyl Line Notation[11] or MDL's RGfile format.[12]

Adapted versions of certain routines from the MAKEBITS program were used to generate "partial fingerprints" on each partial structure, and modified versions of the TreeTrace and Bubble-Up algorithms were implemented to accumulate these to form one fingerprint for each compound covered. Those fragments contained entirely within a single partial structure ("intra-PS" fragments) are assigned to a partial fingerprint bitstring associated with that partial structure, and the modified Bubble-Up algorithm logically ORs together the partial fingerprints for each partial structure combination required. Fragments spanning two or more partial structures ("inter-PS" fragments) are assembled from partial fragments during the Bubble-Up process itself and added to the partial fingerprint for the highest-level partial structure with which they are associated.

In the initial results described here, only "augmented atom" fragments (an atom with its immediate neighbours and their connecting bonds) are generated; work is currently in hand to extend this to the full range of fragment types available in the standard MAKEBITS program.[10]

RESULTS

Data from demonstration files provided with Daylight software were used to generate fingerprint files by both the enumeration and Markush routes. The PAM95 library contains 600 small molecules, the PEPTIDE95 library contains 1280 peptides, and the PEPTOID95 library contains 1280 peptoids (N-substituted glycines). For each CL, the fingerprint files generated by the two routes were identical (after sorting to deal with different enumeration orders), and the program timings given in Table 1 show that the Markush method is significantly faster (though no attempt was made to optimize the program code for speed in either method) and clearly avoids the need to store the large enumerated connection table files.

DISCUSSION

The Markush method for fingerprint generation clearly provides significant time savings, which are likely to be

Table 2. File Sizes (Kilobytes) for the Three Sample CLs

library	CHORTLES and monomer tables (TDT files)	enumerated connection tables	fingerprints (BCI format files)
PAM95	7.5	1200	57
PEPTIDE95	7.5	2500	100
PEPTOID95	7.6	2400	106

greater for larger CLs, as the ratio of number of compounds to number of building blocks (partial structures in the EFCL) increases. The fragment Bubble-Up principle is also applicable to any structural descriptor which is essentially an additive property of its component parts, such as many topological indexes.

Despite the advantage of more rapid generation of structural descriptors, the rate-limiting step for diversity analysis of a CL by clustering its compounds is likely to be the pairwise comparison of the fingerprints (for which the time requirements are related to the square of the number of compounds). However, the present demonstration of the applicability of Markush structure-handling techniques to diversity analysis in CLs also provides a basis for further work. For example, the Bubble-Up algorithm could be adapted for direct generation of a “modal fingerprint”,[13] characteristic of a CL as a whole, in which each bit is set if the corresponding fragment appears in more than a certain proportion of the compounds covered. For clustering purposes, in which pairwise similarities between the compounds are required, these could be calculated “on the fly” without the need to output a file of fingerprints. With a clustering method such as the popular Jarvis−Patrick nonhierarchical method, only the top 20 or so nearest neighbors of each compound are required, and upperbound calculations might be used in order to avoid full calculation of similarities which could not occur in the top 20 list. A similar approach might also be used to find reciprocal nearest neighbors, as required in some of the hierarchical clustering methods which have recently been found to perform significantly better than the Jarvis−Patrick method.[14]

ACKNOWLEDGMENT

Thanks are due to John Bradshaw (GlaxoWellcome), Rob Brown, Bev Eccles, and Yvonne Martin (Abbott Laboratories) for helpful discussions and financial support of this work. Also to Craig James, Yosi Taitz, and Jeremy Yang (Daylight Chemical Information Systems Inc.) for provision of and assistance with Daylight software; to Valerie Gillet and Peter Willett (Sheffield University) for assistance with computing facilities; and to Gary Wiggins (Indiana University) where J.M.B. spent two months as a Visiting Scholar in Fall 1995 and gained experience in the use of the Daylight toolkits.

REFERENCES AND NOTES

(1) Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. Measuring diversity: experimental design of combinatorial libraries for drug discovery. *J. Med. Chem.* **1995**, *38*, 1431−1436.

(2) Lynch, M. F.; Holliday, J. D. The Sheffield generic structures project - a retrospective review. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 93−936.

(3) Barnard, J. M. A comparison of different approaches to Markush structure handling. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 64−68.

(4) Meyer, E.; Schilling, P.; Sens, E. Experiences with input, translation and search in files containing Markush formulae. In *Computer Handling of Generic Chemical Structures*; Barnard, J. M., Ed.; Aldershot: Gower, 1994; pp 83−95.

(5) Barnard, J. M.; Downs, G. M. Applications of Markush structure techniques to handling combinatorial libraries. Presented at a symposium organised by the Division of Chemical Information at the 210th National Meeting of the American Chemical Society, Chicago, IL, Auguest 20−24, 1995.

(6) Barnard, J. M.; Lynch, M. F.; Welford, S. M. Computer storage and retrieval of generic structures in chemical patents. Part 4. An extended connection table representation for generic structures. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 160−164.

(7) Downs, G. M.; Gillet, V. J.; Holliday, J. D.; Lynch, M. F. Computer storage and retrieval of generic chemical structures in patents. 10. The generation and logical bubble-up of ring screens for structurally-explicit generics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 215−224.

(8) Daylight Chemical Information Systems, Inc., 27401 Los Altos, Suite #370, Mission Viejo, CA 92691, USA. http://www.daylight.com.

(9) Siani, M.; Weininger, D.; James, C. A.; Blaney, J. M. CHORTLES: a method for representing oligomeric and template-based mixtures. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1026−1033.

(10) Barnard, J. M.; Downs, G. M. Chemical fragment generation and clustering software. *J. Chem. Inf. Comput. Sci.* In press.

(11) Ash, S. SYBYL Line Notation: Full Markush, combinatorial and query specification in a single language. *J. Chem. Inf. Comput. Sci.* Submitted for publication.

(12) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244−255.

(13) Shemetulskis, N. E.; Weininger, D.; Blankley, C. J.; Yang, J. J.; Humblet, C. Stigmata: an algorithm to determine structural commonalties in diverse datasets. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 862−871.

(14) Brown, R. D.; Martin, Y. C. Use of structure−activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.

CI960091C