

# RECAP—Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry

Xiao Qing Lewell,\* Duncan B. Judd, Stephen P. Watson, and Michael M. Hann

Glaxo Wellcome Research and Development, Medicines Research Centre, Gunnels Wood Road, Stevenage, Hertfordshire SG1 2NY UK

Received August 15, 1997

The use of combinatorial chemistry for the generation of new lead molecules is now a well established strategy in the drug discovery process. Central to the use of combinatorial chemistry is the design and availability of high quality building blocks which are likely to afford hits from the libraries that they generate. Herein we describe “RECAP” (Retrosynthetic Combinatorial Analysis Procedure), a new computational technique designed to address this building block issue. RECAP electronically fragments molecules based on chemical knowledge. When applied to databases of biologically active molecules this allows the identification of building block fragments rich in biologically recognized elements and privileged motifs and structures. This allows the design of building blocks and the synthesis of libraries rich in biological motifs. Application of RECAP to the Derwent World Drug Index (WDI) and the molecular fragments/building blocks that this generates are discussed. We also describe a WDI fragment knowledge base which we have built which stores the drug motifs and mention its potential application in structure based drug design programs.

## INTRODUCTION

The pharmaceutical industry has seen a major change in the way in which lead molecules are identified. An effective combination of automation, high throughput screening, and combinatorial chemistry now allows the screening of libraries of vast numbers of compounds against prospective biological targets to identify new lead molecules. The design of combinatorial libraries falls into two broad classes: (i) those which are designed to express maximum diversity for essentially random screening against a broad spectrum of targets and (ii) those that are biased toward specific targets or classes of target.<sup>1</sup> In designing diversity based libraries diversity selection algorithms are usually employed.<sup>2</sup> These algorithms either select in building block space or product space where a large number of descriptors can be computed for the prospective molecules. However in designing libraries biased for certain targets, the approach is usually undertaken where structural motifs suitable for the target are identified and incorporated into the libraries, typically, through the building blocks. The screening of diversity based libraries is based on the theoretical assumption that the greater the diversity of a collection of molecules the greater the chance of certain molecules hitting some biological targets. In a target biased library, the incorporation of appropriate structural motifs should enhance the probability of identifying molecules active against that specific target.

Several techniques have been developed to identify structural motifs/fragments common to molecules which interact with specific biological targets or target classes.

Pharmacophore modeling techniques<sup>3</sup> have been developed where, for a given set of ligands, the importance of hydrophobic, hydrophilic, and charged functional groups and their geometric relationships on biological activity can be deduced. The deduced model usually loses any substructural information associated with the ligand structures since they rely on a set of “pharmacophoric points” rather than atom connectivities. Alternatively, if the structure of a protein and its mechanism of interaction is known, then the complementary interacting functional groups can be identified and incorporated into the ligand design.<sup>4</sup> Another, and perhaps more ad hoc, approach is to identify substructural motifs in a set of active ligand structures. The structures can be scanned by eye, if the number of ligands is modest, or in the case of larger numbers grouped into clusters by cluster analysis techniques and common fragments identified. The problem with this approach is that the interaction motifs can be biased by an investigator’s opinion, and if the set of structures is particularly diverse, then common substructures embedded in the diverse set may not be recognized.

More systematic procedures for identifying fragments or substructures are exemplified by the work of Klopman,<sup>5</sup> where he devised an algorithm to break down a structure in a systematic fashion to obtain fragments containing between 3 and 12 atoms. Indeed substructures obtained were then related to biological activities to identify “biophores” which he concluded were mainly around 8–10 atoms. Muskal<sup>6</sup> also developed the “Theraprints” concept where drug molecules were broken into energetically stable fragments, and these “Theraprints” were then used for reagent selection and structure–activity recognition. Related in concept is Fujita’s EMIL<sup>7</sup> program. He developed a computer-aided system

\* To whom all correspondence should be addressed. E-mail: xql4406@ggr.co.uk Fax: +44(0)1438 764918.

to allow bioisosteric structural transformation where, given a biologically active molecule, structural bioisosteric groups can be searched in a database of agrochemical agents.

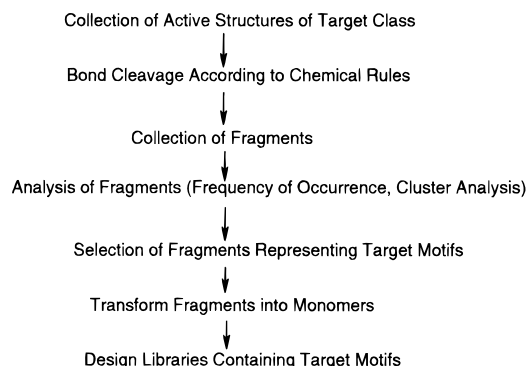
Other published work relating to fragmentation involves retrosynthetic analysis and synthesis planning programs such as LHASA,<sup>8</sup> CAESA,<sup>9</sup> EROS,<sup>10</sup> and the toxicity prediction program DEREK.<sup>11</sup> All these systems rely on some form of knowledge base where fragments are built either by manually extracting experts' knowledge or by theoretically computing bond energies.

In a previous paper,<sup>12</sup> we described a method for identifying drug-like building blocks based on substructural similarity to a drug collection such as the Derwent World Drug Index (WDI).<sup>13</sup> This approach was based on selecting a diverse set of potential building blocks and computing the structural similarity of each selected building block to the WDI molecules. The building blocks that showed greatest similarity to a substructural part of a WDI molecule were then selected for library synthesis. However, as discussed in the paper, there is an inherent drawback to this method. Because the substructural match is based on the match of fingerprints of atom paths in a molecule, the atom path or substructure identified may be embedded in a larger molecular framework. For example, phenylethylamine and naphthylethylamine both contain the phenylethylamine substructure. Using the previous substructural similarity approach, the two molecules will both be selected even though it may only be desired to find the phenyl analogue. This may result in the potential mismatch of true substructures. Also a prerequisite of the previous approach is the prior knowledge of building block structures in order to compute the substructural similarities to drug molecules. These building blocks are usually selected from the available chemicals, and therefore there is no design element to the process.

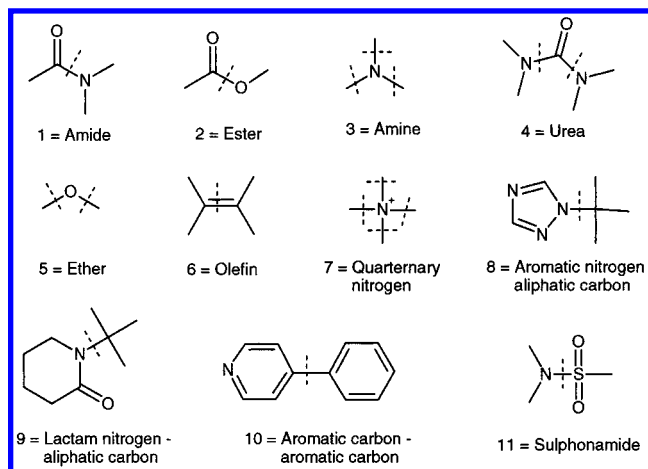
In contrast to our previous approach, we herein describe a new technique for identifying common motifs/fragments in biologically active molecules, based on fragmenting molecules around bonds which are formed by common chemical reactions. Motifs/fragments obtained by this technique are always true substructures, and thus the problem associated with our previous method is avoided. Because we have fragmented molecules into fragments, we can identify those fragments that are likely to be responsible for biological activities. We can therefore design building blocks containing these biological fragments and not restricted by what is available by chemical suppliers. In the process of fragmentation, we have retained the knowledge of the chemical environments of the fragments by assigning them "isotopic labels" to represent the classes of bonds in the prefragmented molecules. The fragments can thus be recombined using known chemistries if desired. The difference between our approach and the numerous approaches discussed above is that we fragment at 11 predefined bond types all of which are amenable to combinatorial chemistry. The resulting fragments are therefore suitable building blocks for combinatorial library synthesis.

## METHOD

**Concept of RECAP.** The RECAP procedure (Figure 1) starts by collecting a set of structures with activity at a common target or target class. Using the set of fragmentation



**Figure 1.** Flow chart of RECAP.

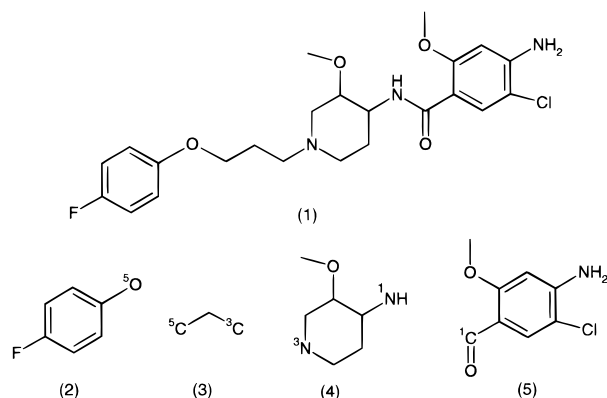


**Figure 2.** Eleven default bond cleavage types.

rules described below each structure is then subject to the RECAP procedure to cleave exhaustively into fragments. These rules are applied such that effectively all susceptible bonds are cleaved in a single pass. Therefore no intermediate structures appear in the final list of fragments. All fragments obtained are collected for subsequent analysis. The analysis includes the frequency of occurrence of each fragment which gives an indication of how often a fragment occurs in the original active structures and cluster analysis which groups similar fragments into a cluster to enable easy identification of fragment patterns. Fragments thus identified can be incorporated as library building blocks for a new design.

**Cleavage Rules.** We have chosen 11 chemical bond types at which to cleave a molecule. These bond types are derived from common chemical reactions. The purpose is therefore to identify fragments which can be synthesized easily from known chemistries. Each molecule is cleaved into fragments if it contains any of the 11 bond types defined in Figure 2. If the terminal fragment to be cleaved contains only small functional groups (hydrogen, methyl, ethyl, propyl, and butyl), the fragment is left uncleaved. Thus in Figure 3 the methoxy group in fragment (4) and (5) is left intact. The reason for this is, first, to avoid generating "uninteresting" small fragments such as methyl in the analysis and, second, with these smaller functional groups attached to the larger fragments such as those shown in Figure 3, more "drug-like" features are retained compared to the "barer" fragments where these small functionalities are cleaved off.

In contrast to Klopman's work, we made the conscious decision to cleave only acyclic bonds so that ring motifs are left intact. This way ring motifs are easily identified



**Figure 3.** A cleavage example of how Cisapride (top) is cleaved into fragments (bottom). ("Isotopic" labels in the fragments denote the chemical environment of the connecting atoms. "Isotopic" types are defined in Figure 2.)

and can be related to a chemist's perception of ring structures.

It should be emphasized that the program also allows the user to define alternative chemical bond types to be cleaved. These alternative bond types could be single or multiple retrosynthetic cleavages suited to any intended specific library or ring bonds where libraries to be synthesized form these ring bonds in situ.

The algorithm is written in C and makes use of the DAYLIGHT toolkit.<sup>14</sup> DAYLIGHT SMARTS notation is used for bond type definitions, and SMILES strings are required for input of molecules.<sup>15</sup>

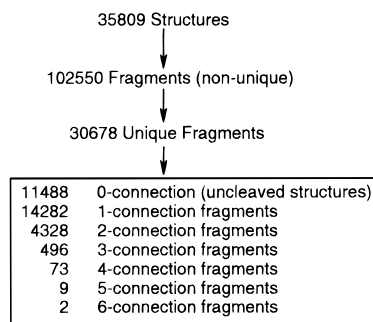
**A Cleavage Example.** An example of a typical cleavage is given in Figure 3. The atoms at the cleavage point in the fragments are given numeric or "isotopic" labels to denote the atom environments they are derived from (Figure 2). Thus <sup>1</sup>C represents an amide carbon and <sup>3</sup>N an amine nitrogen, Figure 3.

The advantage of retaining knowledge about the chemical environments of the fragments is to aid synthesis planning. Thus only the appropriate fragments are incorporated into the design of monomers to react with appropriate chemistries. The number of attachment points per fragment may also influence the design process. Thus a fragment can be considered for a terminal monomer if it has one connection point (1-connection fragments) or as a core template if it has more than one connection point ( $\geq 2$ -connection fragments). For example, on RECAP of molecule (1) in Figure 3 fragments (2) and (5) could be considered as monomers whereas (3) and (4) as core templates.

## RESULTS

We now illustrate the RECAP technique with some application examples. We will discuss some results on fragmenting the Derwent World Drug Index and illustrate the process of identifying building blocks for library design.

**Gross Observations on Fragmenting World Drug Index.** The 1995 version of the World Drug Index<sup>13</sup> under DAYLIGHT Chemical Information Systems (WDI954)<sup>14</sup> contains approximately 50K documented molecules of pharmaceutical interest. We have selected all molecules with reported biological activities. The resulting set of 35K molecules was then fragmented using the RECAP procedure. Figure 4 shows the numbers of fragments obtained. The



**Figure 4.** Numbers of fragments obtained after RECAP of WDI structures.

frequency of occurrence of these fragments ranges between ca. 1500 times for dimethylamino fragment (C[3N]C)<sup>15</sup> to many fragments which occur only once. The decline of fragment frequency is very rapid. For visual inspection, we have plotted the logarithmic scale of the frequency distribution in Figure 5.

One interesting observation is that as the number of structures increases in a database, the number of unique fragments obtained saturates toward approximately 2/3 of the database size (given the 11 bond type definitions we have used for cleavage). This can be seen from Table 1, where we have fragmented different proportions of the WDI database, two corporate databases GLAX and WELREG, and the commercial database SPRESI.<sup>16</sup>

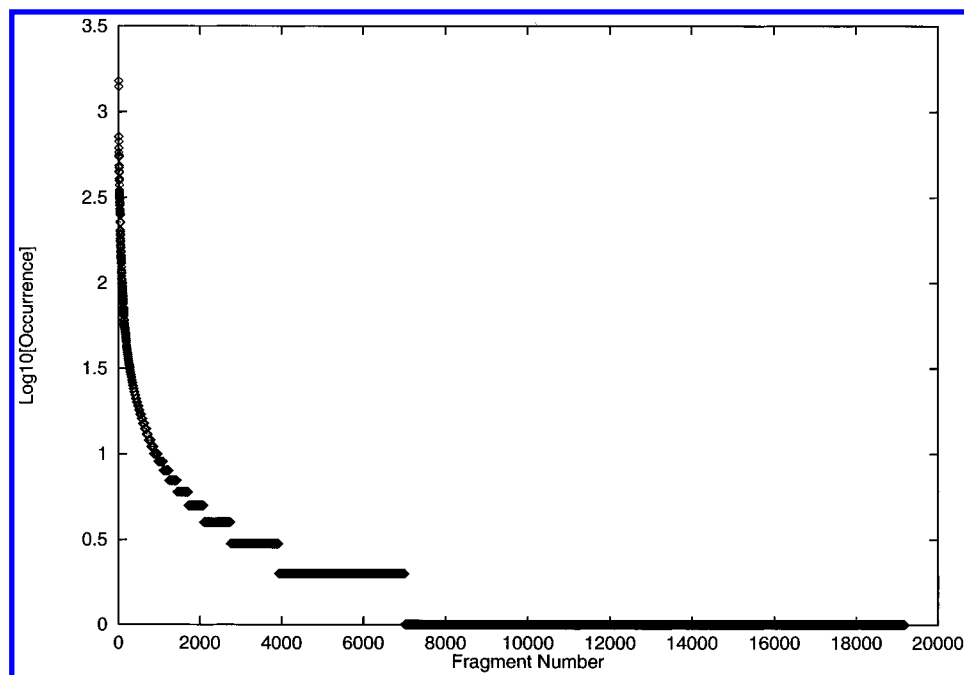
Since ring bonds are not cleaved, the number of fragments obtained per molecule is proportional to the number of bonds cleaved (no. of fragments = no. of bonds cleaved + 1). However, since each molecule is unique in its bond composition, there is little correlation between the size of the original structure and the number of fragments generated per structure. This can be seen from Figure 6, where Spearman's<sup>17</sup> rank correlation coefficient ( $r$ ) is 0.5 between the number of fragments generated and the molecular weight of the original structure before cleavage.

The molecular weight distributions for structures which have been fragmented and for fragments generated are shown in Figure 7. [N.B. 0-connection fragments, i.e., intact structures are excluded. Thus from the 30K fragments including 0-connection fragments obtained on cleaving WDI structures, this resulted in approximately 20 000 fragments, see Figure 4. These 20K fragments are used for further studies, and any subsequent discussions in the text refer to this set.] There is a considerable shift of the distribution peak toward lower molecular weight for fragments compared to prefragmented structures. The significance of this will be discussed in the following section on applications to combinatorial chemistries.

Some examples of 1- through 6-connection fragments are shown in Figure 8.

## APPLICATIONS OF RECAP

**Drug Motif Extraction on World Drug Index.** An example application of RECAP is to identify molecular fragments which are likely to be associated with specific biological activities. In the set of 35K WDI structures discussed in the previous section, certain biological activity keywords are associated with each structure. The WDI



**Figure 5.** Frequency distribution of WDI fragments.

activity labels are either single or multiple activity keywords as shown by the examples below:

<chem>C[N+](C)(C)CCO</chem> <sup>15</sup>	PARASYMPATHOMIMETICS;VITAMINS-B
<chem>O=Cc1cccn1</chem>	ANTIINFLAMMATORIES
<chem>NC1CC1c2cccc(F)c2</chem>	PSYCHOSTIMULANTS;MAO-INHIBITORS; ANTIDEPRESSANTS

We have taken the approach to duplicate the structures with multiple activities such that a structure is duplicated and associated with only one activity:

<chem>C[N+](C)(C)CCO</chem>	PARASYMPATHOMIMETICS
<chem>C[N+](C)(C)CCO</chem>	VITAMINS-B
<chem>O=Cc1cccn1</chem>	ANTIINFLAMMATORIES
<chem>NC1CC1c2cccc(F)c2</chem>	PSYCHOSTIMULANTS
<chem>NC1CC1c2cccc(F)c2</chem>	MAO-INHIBITORS
<chem>NC1CC1c2cccc(F)c2</chem>	ANTIDEPRESSANTS

This way we can fragment molecules, and for each fragment obtained, its occurrence for a single biological activity type can be counted.

Figure 9 shows some examples of the one-connection fragments generated and the top occurring therapeutic class for each fragment. The first number denotes the total number of occurrence of the fragment within the WDI molecules, whereas the second number denotes the number of occurrence of the fragment within the labeled therapeutic class. Clearly the total number of occurrences of the fragment and its occurrence within a therapeutic class together give an indication of whether a fragment is specific for that class or whether it is generic, i.e., occurs in many unrelated therapeutic classes.

**Table 1.** Number of Unique Fragments Obtained on Fragmenting Different Sized Databases

database	no. of structures	no. of unique fragments	ratio (fragment/structures)
WDI	3750	4905	1.31
WDI	7500	8745	1.17
WDI	11250	12362	1.10
WDI	15000	15674	1.04
WDI	18750	18753	1.00
WDI	22500	21836	0.97
WDI	26250	24813	0.95
WDI	30000	27677	0.92
WDI	33750	30507	0.90
WDI	35809	30678	0.86
WELREG	121753	78146	0.64
GLAX	312298	201205	0.64
SPRESI	1808369	1191798	0.66

For example, if we take the 4th, 9th, and 20th fragments in Figure 9, they have top occurring therapeutic classes as “SYMPATHOLYTICS–BETA”, “ANTIBIOTICS”, and “LEUKOTRIENE-ANTAGONISTS”, respectively. The “SYMPATHOLYTICS–BETA” fragment is quite generic, whereas the “ANTIBIOTICS” fragment is very specific, and the “LEUKOTRIENE-ANTAGONISTS” fragment is somewhere in between as reflected by the numbers and natures of different therapeutic classes for each fragment, Figure 10. It is also interesting to note that several top occurring therapeutic keywords for the “SYMPATHOLYTICS–BETA” fragment are related, and most of the classes for “LEUKOTRIENE-ANTAGONISTS” fragment are mainly antiinflammatory in nature [See notes in Figure 10.].

The advantage of the RECAP analysis is that it can identify common fragments within a therapeutic class that would be difficult to find with conventional structural based clustering methods, e.g., when the common fragment occurs in a diverse set of structures having different molecular connectivities. To illustrate this point, there are 344 LEUKOTRIENE-ANTAGONISTS in the WDI database which are of a variety of structural types. Using the conventional Ward<sup>18</sup> clustering method based on the distance matrix derived from the

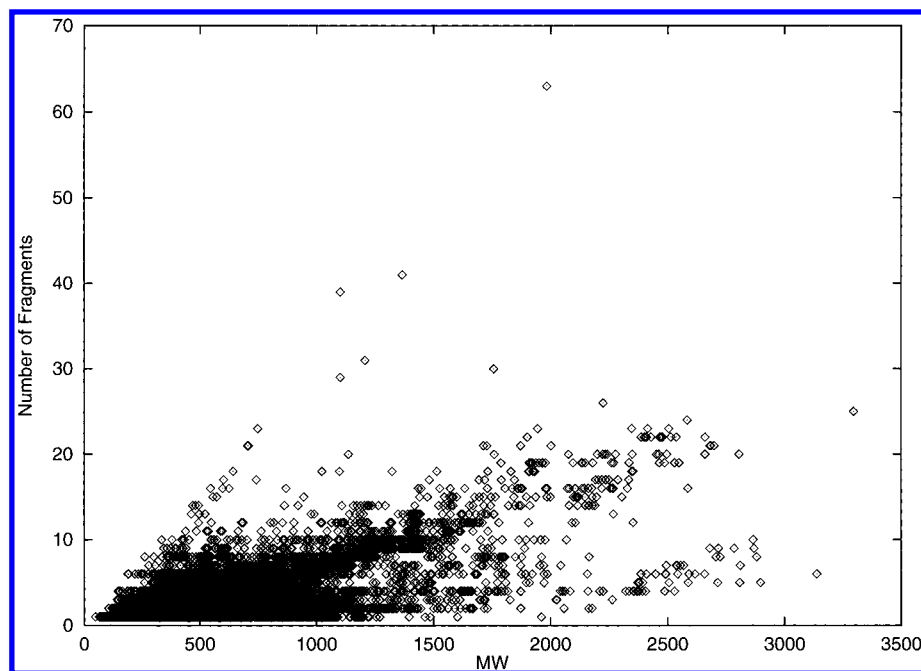


Figure 6. Number of fragments generated per structure versus MW of structure.

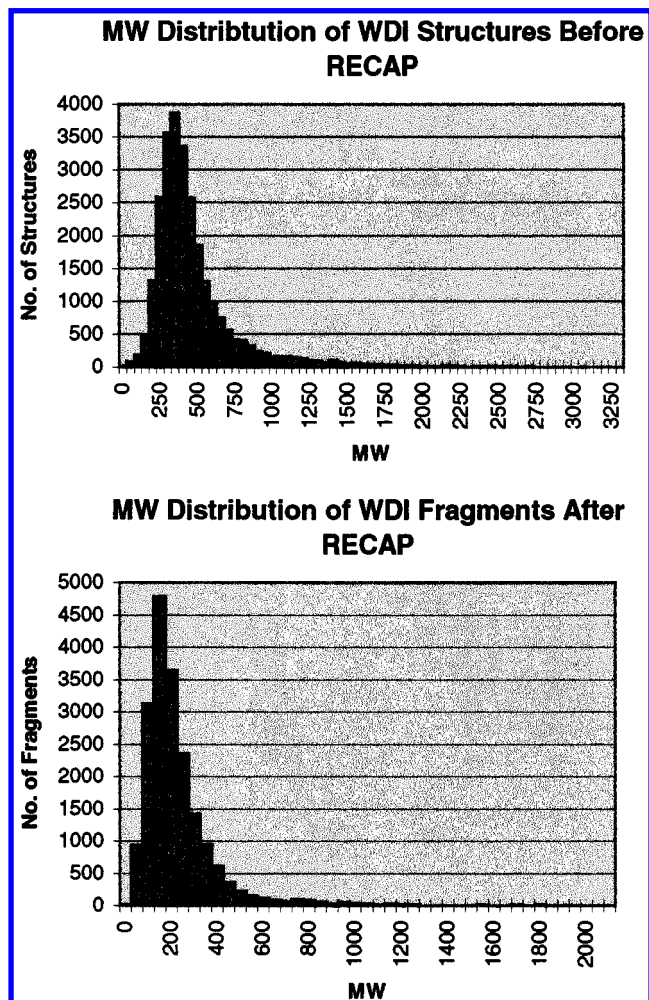


Figure 7. Molecular weight distributions of WDI structures and fragments.

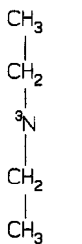
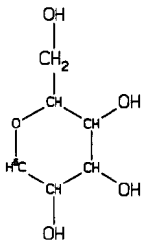
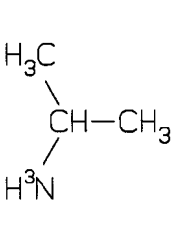
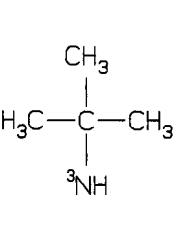
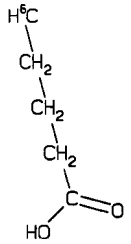
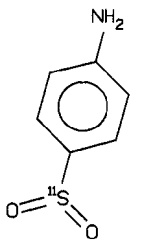
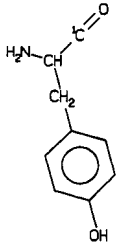
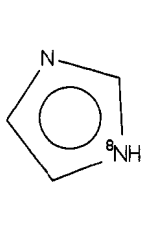
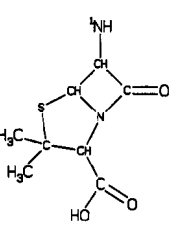
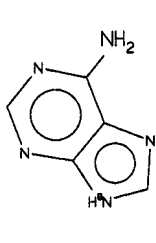
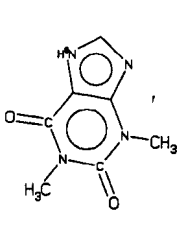
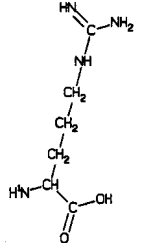
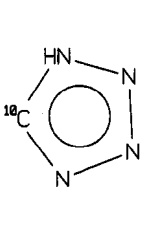
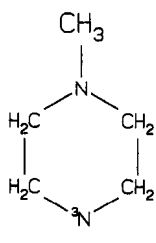
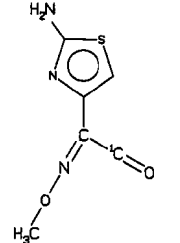
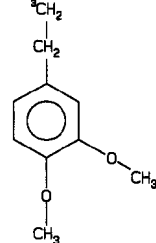
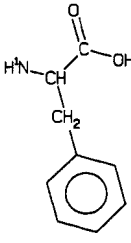
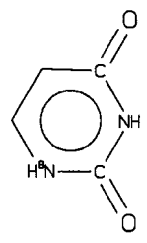
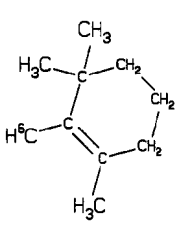
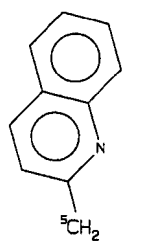
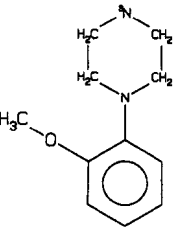
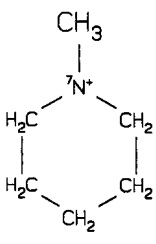
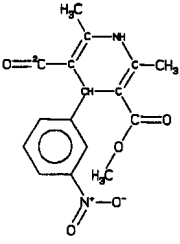
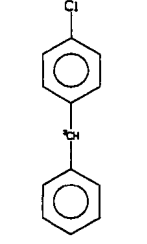
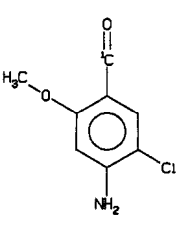
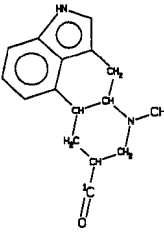
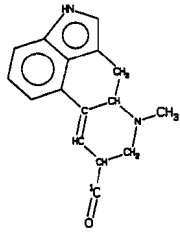
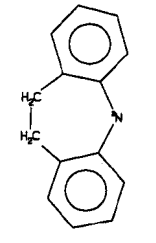
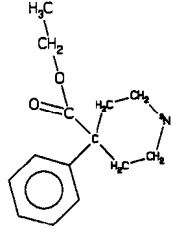
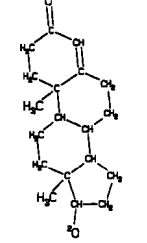
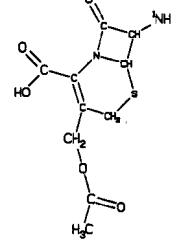
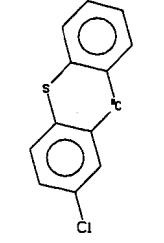
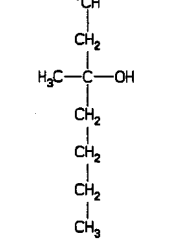
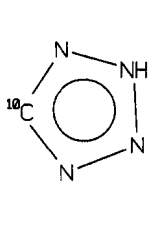
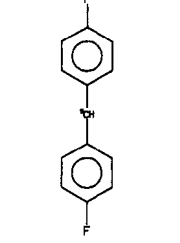
Daylight fingerprint of molecules, 344 molecules can be grouped into several clusters. Although one of the clusters contains 45 of the 49 2-oxomethylene quinoline analogues, it also contains 25 nonquinoline structures. In addition, two

other clusters also contain the remaining 2-oxomethylene quinolines analogues [Supporting Information is available for clusters of 344 leukotriene antagonists.] For the 49 structures containing 2-oxomethylene quinoline, there are enough structural variations that the importance of the methylene quinoline motif in leukotriene antagonists may not be obvious, Figure 11. However, using our RECAP analysis, we have immediately identified among the WDI structures that 2-oxomethylene quinoline is a potential leukotriene motif since it occurs 47 times in this class compared to a total occurrence of 55 times in all of the WDI structures. In fact, the 2-quinolylmethoxy phenyl motif as an essential structural fragment for a large proportion of the leukotriene synthesis inhibitors has been discussed in the past.<sup>19</sup> Our finding using the RECAP data mining tool clearly supports the conclusion derived from the traditional medicinal chemistry analysis on leukotriene structures.

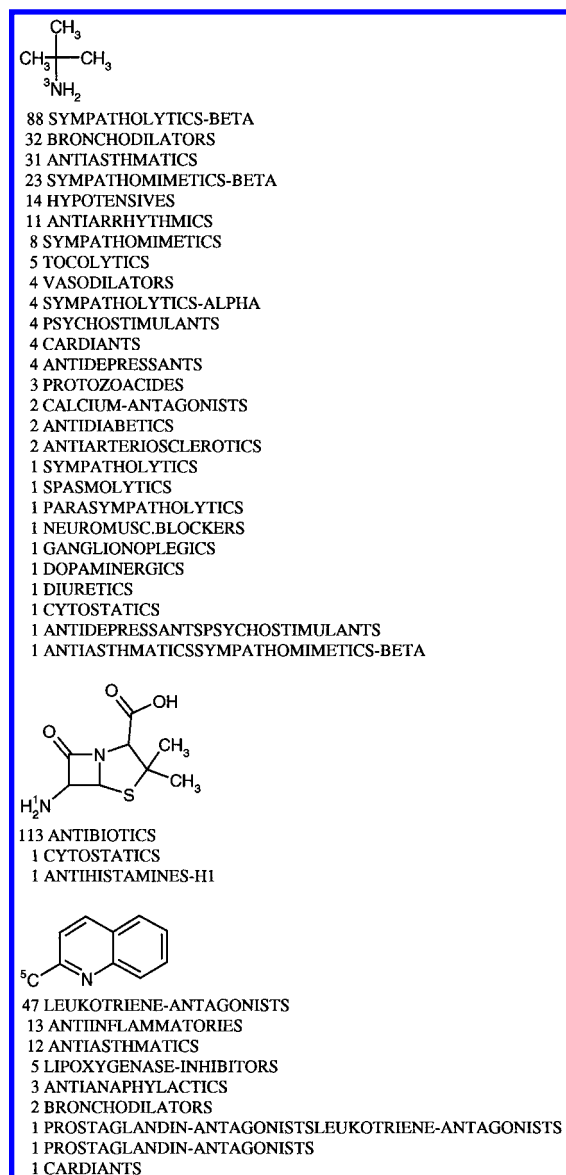
It is worth contrasting RECAP with Stigmata<sup>20</sup> as a technique for finding substructural motifs. Stigmata is designed to find structural commonalities in a diverse dataset based on the concept of a modal fingerprint. In Stigmata, structures are not cleaved, and common structural paths are identified by matching overlapping fingerprints. Without prior knowledge about the quinoline motif, it is difficult to identify it in the diverse set of 344 leukotriene antagonists without having to go through a trial and error process of setting thresholds.<sup>20</sup> However the advantage of having adjustable threshold ratio within Stigmata is to identify different subsets of common substructures, although the substructures identified may not form readily retrosynthesizable building blocks. RECAP, on the other hand, cleaves the structures into fragments based on retrosynthetic chemistry. The fragments generated are therefore limited by the retrosynthetic bond type definitions. However the advantage is that once fragments are identified to be prerequisites for activity, they can be readily transformed into chemical building blocks for synthetic programs. Methods such as RECAP, Stigmata, and conventional structure based cluster-





714 213 PSYCHOSEDATIVES 	473 112 ANTIBIOTICS 	254 121 SYMPATHOLYTICS-BETA 	181 88 SYMPATHOLYTICS-BETA 	154 119 PROSTAGLANDINS 	141 126 ANTISEPTICS 
130 99 OPIOIDS 	124 46 FUNGICIDES 	116 113 ANTIBIOTICS 	114 45 VIRUCIDES 	99 30 CARDIANTS 	97 34 BRADYKININ-ANTAGONISTS 
95 88 ANGIOTENSIN-ANTAGONISTS 	94 40 NEUROLEPTICS 	92 87 ANTIBIOTICS 	65 32 CALCIUM-ANTAGONISTS 	68 11 ANGIOTENSIN-AGONISTS 	56 35 ANTIBIOTICS 
55 51 VITAMINS-A 	55 47 LEUKOTRIENE-ANTAGONISTS 	51 16 SYMPATHOLYTICS-ALPHA 	46 23 NEUROMUSCULAR BLOCKERS 	36 33 CALCIUM-ANTAGONISTS 	26 13 ANTIHISTAMINES-H1 
23 10 ANTIEMETICS 	21 16 ANTISEROTONINS 	20 9 DOPAMINERGICS 	20 20 PSYCHOSTIMULANTS 	19 16 ANALGESICS 	19 18 ANDROGENS 
19 19 ANTIBIOTICS 	18 12 PSYCHOSEDATIVES 	17 11 PROSTAGLANDINS 	17 16 ANGIOTENSIN-ANTAGONISTS 	16 12 DOPAMINERGICS 	

**Figure 9.** Examples of some WDI fragments with their top therapeutic classes. The first number in a header box denotes the total occurrence of the fragment within WDI collection, whereas the second denotes the occurrence within the labeled therapeutic class.



**Figure 10.** Three examples of WDI fragments with their occurrences in therapeutic classes. Note that not all classes are independent of each other, i.e., "ANTIINFLAMMATORIES" may co-occur with "LEUKOTRIENE-ANTAGONISTS" as activity label for some WDI molecules (see text).

weight for WDI structures (MW = 450) and for fragments (MW = 150). Since a typical library is usually made from two or three building blocks including core and monomers, when two or three fragments are joined together to form a library product, the expected molecular weight distribution for the library would be centered around MW = 300–450, the desired molecular weight profile for many drug-like molecules.

To identify fragments that can be considered as potential monomers, fragments are grouped into clusters according to their structural similarity. We have used the Jarvis-Patrick algorithm<sup>21</sup> under DAYLIGHT.

Figure 12 shows an example of a cluster of indole-containing 1-connection fragments. A "monomer motif" is defined as either a frequently occurring fragment or a representative fragment of a cluster where the individual cluster members occur less frequently on their own. This representative fragment could simply be that selected by a

chemist which is amenable to chemistry. The monomer motif is then translated into the appropriate monomer. Thus in the case of the indole fragments derived from amide bond formation, Figure 12, they are transformed into amines which are able to undergo, for example, reductive amination. Obviously some protective group strategy of reactive functional groups may be required if the desired monomers contain any competing reactive functionalities. Once the potential monomer is identified, its availability is searched for using databases of available molecules. If it is not available, the potential monomer can be substituted by the most similar available monomer (Figure 13) or, if desirable, can be synthesized. In this way, the library designed will incorporate known target motifs and therefore is expected to have higher success rates in generating leads.

It is worth contrasting the way we identify library building blocks in RECAP with the work published by Bemis and Murcko.<sup>22</sup> In their paper, they have analyzed drug molecules according to their shapes. With the aid of cluster analysis, they have arrived at some common molecular frameworks (ring templates with attachment positions to attach substituents and chain length of up to nine atoms long) to represent majority of the shapes for drug molecules. Their work gives an indication of molecular skeletons for designing drug molecules. However because the structural information is condensed to molecular framework, the atomic composition of the framework is lost, making design more difficult. In contrast, RECAP keeps all atomic information for the building blocks identified. It is therefore easier to use known chemistries that produced these drug molecules to recombine the specific building blocks identified into new molecules.

#### Database Building of WDI Fragment Knowledge Base.

Since we have fragmented the WDI structures and obtained information relating to fragment occurrence within therapeutic classes, this information would be very valuable for researchers to interrogate. We have built a DAYLIGHT THOR database which holds data items shown in Table 2 for each of the 20K fragments obtained. Normal search capabilities within a DAYLIGHT database are available. For example, structure/substructure and similarity searches on fragments, keyword searches to identify fragments associated with specific biological classes, and numeric searches to obtain frequency of occurrence of fragments within biological classes can be easily carried out. The number of connection points of fragments is also stored to enable easy identification of potential core or monomer fragments for library designs. Figure 14 shows a snapshot of the WDI fragment knowledge base we have built.

#### CONCLUSIONS

The RECAP technique which we have developed is a powerful tool for identifying biologically privileged structural motifs and fragments for use in the synthesis of combinatorial libraries. The RECAP technique cleaves molecules at bonds amenable to combinatorial chemistry. The fragments and motifs can therefore be readily used as building blocks to prepare combinatorial libraries rich in biologically privileged substructural motifs. We believe that the RECAP technique will prove particularly powerful in the design and synthesis of libraries focused on specific biological targets. It should be emphasized that the RECAP technique can only identify



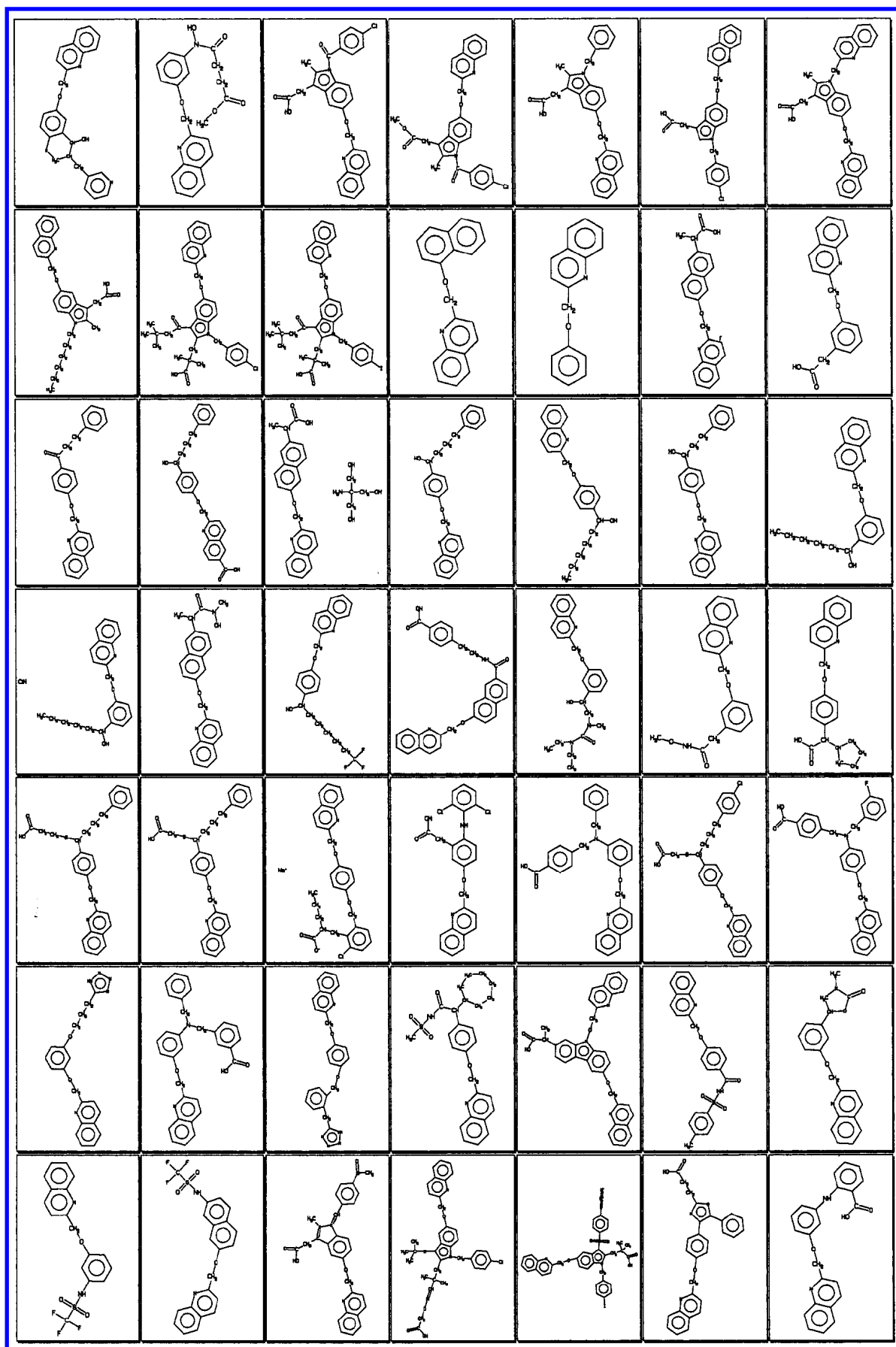
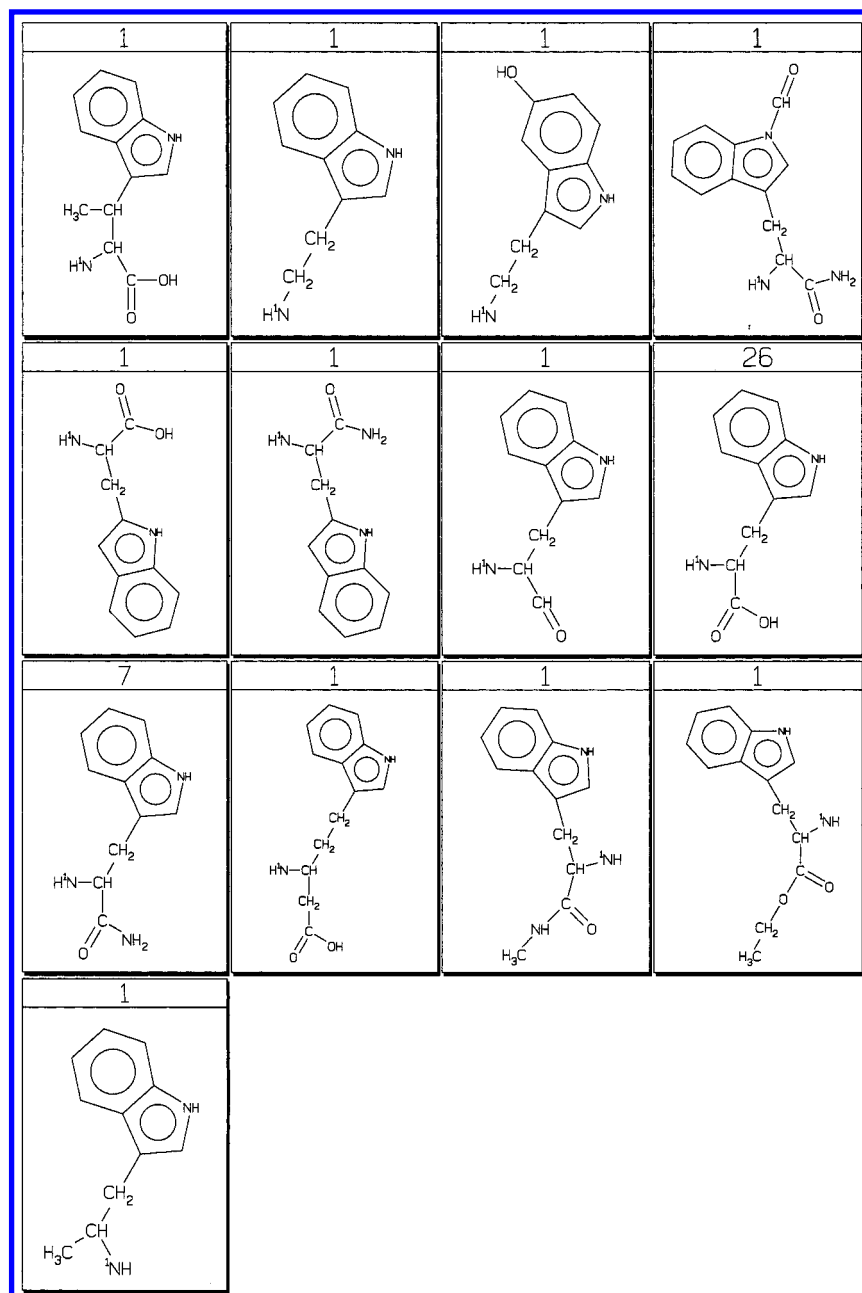
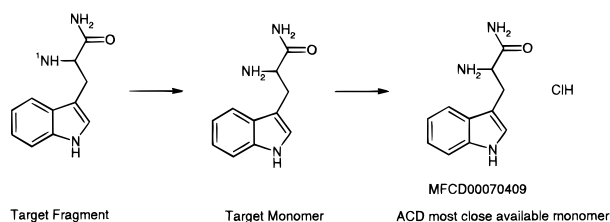


Figure 11. Quinoline-containing leukotriene antagonists.



**Figure 12.** A cluster of indole-containing fragments. "Isotopic" label "1" denotes the nitrogen came from an amide environment. Header box numbers denote occurrence in the WDI collection



**Figure 13.** Transforming target fragment to a target monomer and search for most close monomers.

fragments from molecules made in the past and may be biased by the number of active analogues in a database. However through different recombination of these structural fragments using combinatorial library technology, it is hoped that new and potentially novel molecules will be produced which will yield higher success rate in lead generation and optimization programs compared to the random approach.

**Table 2.** WDI Fragment Knowledge Base Data Types

data type	meaning
SMILES	represents the chemical structure of a fragment
FREQ	represents total number of occurrences of the fragment within the subset of WDI having biological activity keywords
CLASS	represents the biological activity class the fragment is associated with
CLASS FREQ	represents the number of occurrences of the fragment within a biological class
CLASS ORDER	represents the orders in which the fragment most frequently occurs in different biological classes
CONNECTION	represents the number of connection points within a fragment in the original uncleaved structure

Other applications of the RECAP technique include data mining of commercial databases and commercial supplies to identify constituent building blocks.

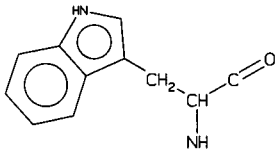
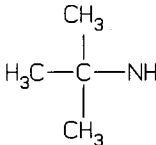
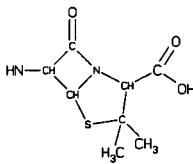
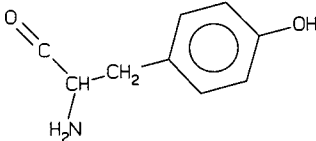
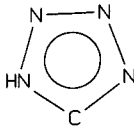
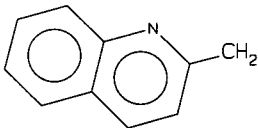
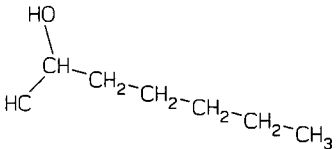
File Hitlist Display Help... Messages...			
Database: <input type="text"/> wdifrag@uk6x02:merlin:xql4406			
Position: 6 Hitlist: 7 Total: 66669			
SMILES	FREQ	CLASS FREQ	CLASS
	338	45	GASTROINTESTHORMONES
	161	88	SYMPATHOLYTICSBETA
	116	113	ANTIBIOTICS
	130	99	OPIOIDS
	95	68	ANGIOTENSINANTAGONISTS
	55	47	LEUKOTRIENEANTAGONISTS
	90	80	PROSTAGLANDINS

Figure 14. WDI Fragment Knowledge Base.

RECAP Analysis of the WDI database and other collections of drug-like molecules would also provide important drug-like fragments for structure based ligand design programs such as SPROUT<sup>23</sup> where a fragment knowledge database is required. Since the fragments have connection points and chemistries associated with them, the de novo ligands generated will be more amenable to chemistry and akin to drug-like molecules.

#### ACKNOWLEDGMENT

We thank all our computational chemistry colleagues who contributed to the discussions, in particular, Drs. D. Butina, J. Bradshaw, R. Carr, D. Green, and A. Leach for their help and input, Mr. Tony Chan for database building, Dr. Y. Mohamed for statistical analysis, and our chemistry colleagues who have taken forward the ideas in their chemical library syntheses. We also thank Prof. Peter Johnson and

his group for our earlier collaboration on CAESA modifications where some of the ideas have been implemented, and Prof. Peter Willet, the referees, and Derwent Information Ltd for comments on the manuscript.

**Supporting Information Available:** Leukotriene antagonists numbering 344 using Ward based clustering model. The label M143 1 means arbitrary model name (M143 = 143rd molecule in the original input list) and for cluster number 1 = cluster 1 (23 pages). See any current masthead page for ordering and Web access instructions.

## REFERENCES AND NOTES

- (1) Young, S. S.; Farnen, M.; Rusinko III, A. Random Versus Rational—Which is better for General Compound Screening, Network Science, Feature 9, Aug 1996. <http://www.awod.com/netsci/Science/Screening/feature09.html>.
- (2) Warr, W. A. Commercial Software Systems for Diversity Analysis, *Perspect. Drug Discovery Des.* **1997**, 7/8, 115–130.
- (3) Leach, A. R. *Molecular Modelling, Principles and Applications*; ISBN 0-582-23933-8, Addison-Wesley Longman Limited: 1996.
- (4) Baggio, R.; Shi, Y.-Q.; Wu, Y.-q.; Abeles, R. H. From Poor Substrates to Good Inhibitors: Design of Inhibitors for Serine and Thiol Proteases. *Biochemistry* **1996**, 35(11), 3351–3.
- (5) Klopman, G. Artificial Intelligence Approach to Structure–Activity Studies. Computer Automated Structure Evaluation of Biological Activity of Organic Molecules. *J. Am. Chem. Soc.* **1984**, 106, 7315–7321.
- (6) Muskal, S. M. Enriching Combinatorial Libraries with Features of Known Drugs, C. Divisions, 20th ACS National Meeting (American Chemical Society, Anaheim, CA, 1995; Vol. 1, p 029).
- (7) Fujita, T. Concept and features of EMIL, a system for lead evolution of bioactive compounds. Trends QSAR Mol. Modell. 92, Proc. Eur. Symp. Struct.-Act. Relat.: QSAR Mol. Modell., 9th 1993, Meeting Date 1992, pp 143–59. Wermuth, C.-G., Eds.; ESCOM: Leiden, The Netherlands, CODEN: 59XTAS. CAN 121: 169406.
- (8) Corey, E. J. Computer-assisted analysis of complex synthetic problems. *Quart. Rev. Chem. Soc.* **1971**, 25(4), 455–82.
- (9) Myatt, G.; Baber, J. C.; Johnson, A. P. New developments in the caesa system for estimation of synthetic accessibility. Book of Abstracts, 21st ACS National Meeting, Chicago, IL, August 20–24 1995; American Chemical Society, Washington, D.C., Issue Pt. 1, COMP-007 CODEN: 61XGAC. AN 1995:920276.
- (10) Gasteiger, J.; Marsili, M.; Hutchings, M. G.; Saller, H.; Loew, P.; Roese, P.; Rafeiner, K. Models for the representation of knowledge about chemical reactions. *J. Chem. Inf. Comput. Sci.* **1990**, 30(4), 467–76.
- (11) Ridings, J. E.; Barratt, M. D.; Cary, R.; Earnshaw, C. G.; Eggington, C. E.; Ellis, M. K.; Judson, P. N.; Langowski, J. J.; Marchant, C. A. Computer prediction of possible toxic action from chemical structure: an update on the DEREK system. *Toxicology* **1996**, 106(1–3), 267–79.
- (12) Lewell, X. Q.; Smith, R. Drug Motif Based Diverse Monomer Selection – Method and Application in Combinatorial Chemistry; *J. Mol. Graphics Model.* **1997**, 15, 43–48.
- (13) Derwent Information Ltd., Derwent House, 14 Great Queen Street, London, WC2B 5DF, UK. Web address: <http://www.derwent.co.uk/>.
- (14) Daylight Chemical Information Systems Inc., 27401 Los Altos, Suite #370, Mission Viejo, CA 92691. Web address: <http://www.daylight.com/>.
- (15) DAYLIGHT SMILES and SMARTS notations to represent chemical structures or fragments. Taking advantage of the isotopic notation of the SMARTS notation, 3 was chosen to represent an amine bond type. [3N] therefore represents an amine nitrogen. For all other isotopic representations, see Figure 2.
- (16) SPRESI93. A chemical substances database extracted by the Academy of Science, USSR and marketed by InfoChem GmbH. A DAYLIGHT version is available from DAYLIGHT.
- (17) Spearman, C. *Am. J. Psych.* **1904**, 15, 88.
- (18) Ward, J. H. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Statistical Assoc.* **1963**, 58, 236–244.
- (19) Matzke, M.; Beckermann, B.; Fruchtmann, R.; Burkhard, F.; Gardiner, P. J.; Goossens, J.; Hatzelmann, A.; Junge, B.; Keldehnich, J. et al. Leukotriene synthesis inhibitors of the quinoline type: parameters for the optimization of efficacy. *Eur. J. Med. Chem.* **1995**, 30(Suppl., Proceedings of the 13th International Symposium on Medicinal Chemistry, 1994), 441s–51s.
- (20) Shemetulskis, N. E.; Weininger, D.; Blankley, C. J.; Yang, J. J.; Humblet, C. Stigmata: An Algorithm To Determine Structural Commonalities in Diverse Datasets. *J. Chem. Inf. Comput. Sci.* **1996**, 36(4), 862–871.
- (21) Jarvis, R. A.; Patrick, E. A. Clustering using a similarity measure based on shared near neighbors. *IEEE Trans Computers*, **1973**, C22, 1025–1034.
- (22) Bemis, G. W.; Murcko, M. A. The properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, 39, 2887–2893.
- (23) Gillet, V. J.; Johnson, A. P.; Mata, P.; Sike, S.; William, P. SPROUT: A Program for Structure Generation. *J. Comput-Aided Mol. Des.* **1993**, 7, 127–153.

CI970429I