ELSEVIER

Drug-motif-based diverse monomer selection: Method and application in combinatorial chemistry

Xiao Qing Lewell and Ryan Smith[1]

Glaxo Wellcome Research and Development, Stevenage, Herts, SG1 2NY, U.K.

[1]Currently in his final year at the University of Surrey, Guildford, Surrey GU2 5XH, U.K.

This article describes a strategy to explore monomer diversity while incorporating drug motif knowledge into the design and selection of monomers for combinatorial chemistry. The process involves collecting from available electronic databases all those molecules that potentially could be monomers. In this manner we have assembled five DAYLIGHT databases, each containing one of the common functional groups: carboxylic acids, aldehydes, nitriles, primary amines, and secondary amines. The molecules in the databases are then subjected to fingerprint and cluster analysis using the Jarvis-Patrick algorithm and profiles of the compounds are calculated relating to molecular weight, H-bond counts, and rotatable bond flexibility. The cluster information and profiles of the molecules are stored back into the databases for similarity and diversity searches, and for profile prescreening of monomers. To apply drug motif knowledge to a selection an application to an aldehyde set is discussed, in which representatives of each cluster in the aldehyde database are compared with drug molecules in the Standard Derwent File (SDF) in one of three ways to select drug motif-based monomers for purchase or synthesis. © *1997 by Elsevier Science Inc.*

Keywords: drug motif, monomer selection, combinatorial chemistry, computational method, database

INTRODUCTION

Within the pharmaceutical industry there is a shift in paradigm in the way new lead compounds are being discovered. This involves both the development of high-throughput screening technologies, which enable the testing of tens of thousands of compounds, and the ability to synthesize large numbers of compounds. Combinatorial chemistry techniques have been developed so that large numbers of compounds can be efficiently produced. The underlying principle of combinatorial chemistry is that a set of building blocks (termed *monomers*) are simultaneously reacted to produce a large number of products. This is illustrated in Figure 1, where 2 + 2 (equals 4) monomers can form 2 × 2 (equals 4) products. Analogously, 100 + 100 (equals 200) monomers will form 100 × 100 (equals 10 000) products.

To explore structural diversity produced by these product molecules, some selection criteria are usually employed to select monomers. Typically descriptors of the monomers are calculated and some reduction and selection methods are used to select a diverse set of monomers on the basis of these descriptors.

Methods based on chemical connectivities such as those developed by Martin et al.[1] and by Weininger[2] based on the Jarvis-Patrick clustering technique,[3] methods based on 3-D conformations such as those developed by Davies and Briant,[4] and methods based on their combinations such as those employed in Tripos software,[5] MSI software,[6] and Pearlman software[7] are designed to give a set of monomers as diverse as possible, on the basis of the assumption that diverse biological activities will be found through sampling chemical diversity. However, experience suggests that certain structural types may be more likely to give biological activity than others. This knowledge is embedded in the Standard Derwent File (SDF).[8]

This article makes the assumption that if those drug motifs occurring in the SDF are incorporated into the monomer design and selection, then the product molecules may be more likely to be biologically active.

OUTLINE OF IDEA

The idea is illustrated in Figure 2. The process involves collecting from electronic databases all those molecules that po-

The Color Plate for this article is on pages 55–56.

Address reprint requests to: Xiao Qing Lewell, Glaxo Wellcome Research and Development, Gunnels Wood Road, Stevenage, Herts, SGI 2NY, U.K.

Paper submitted to Electronic Conference of the Molecular Graphics and Modelling Society, October 1996

Journal of Molecular Graphics and Modelling 15, 43–48, 1997
© 1997 by Elsevier Science Inc.
655 Avenue of the Americas, New York, NY 10010

1093-3263/97/$17.00
PII S1093-3263(97)00009-X

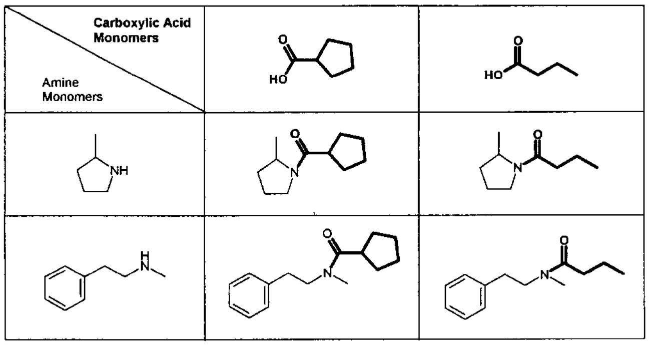


Figure 1. Illustration of a 2 × 2 combinatorial library using acid and amine monomers as reactants.

tentially could be monomers. The monomers are subjected to cluster analysis, and a representative (termed the *centroid*) from each cluster is selected. Each of the centroid monomers is compared with the 30 000 SDF molecules in several ways:

1. Whole-molecule comparisons: in which the monomer as compared to each of the SDF molecules and the highest similarity score is recorded.
2. Exact substructure occurrence of the monomer in SDF molecules: in which an exact substructure search is carried out and the number of occurrences recorded.
3. Similar substructure occurrence of the monomer in SDF molecules: in which part of an SDF molecule is similar to the monomer as measured by a similarity score (defined below). The occurrence of similar substructures is recorded.

The whole-monomer/whole-molecule similarity measure in option 1 will give monomers that are close in size to the SDF molecules. However, since the library products usually are formed from at least two monomers, it is more meaningful to detect substructures contained in SDF molecules that are similar to the monomer in question, so that the product molecules formed will contain these substructural (or “druglike”) motifs.

Thse centroid monomers that are either highly similar as determined by analysis 1, or occur frequently as determined by analysis 2 or 3, are then selected to initiate chemistry. In this way, diversity is covered by these centroid monomers as the starting set and similarity criteria will select from this diverse set monomers that are also “druglike” compared to SDF molecules.

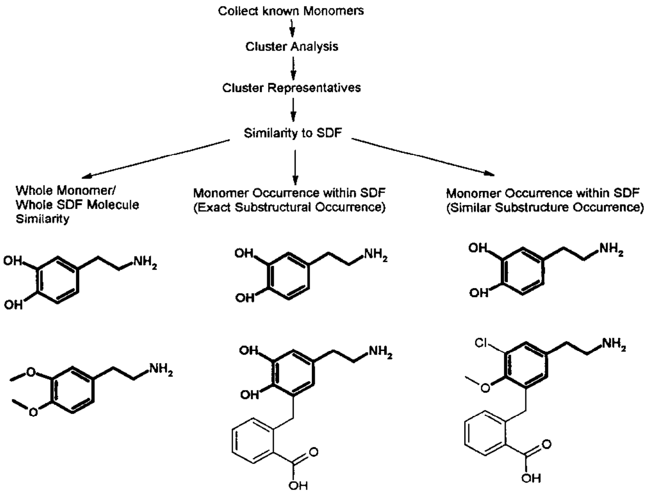

Collect known Monomers
Cluster Analysis
Cluster Representatives
Similarity to SDF
Whole Monomer/ Whole SDF Molecule Similarity
Monomer Occurrence within SDF (Exact Substructural Occurrence)
Monomer Occurrence within SDF (Similar Substructure Occurrence)
OH
OH
NH₂
O
O
NH₂
OH
OH
NH₂
OH
OH
NH₂
O
OH
OH
OH
NH₂
Cl
O
NH₂
O
OH

Figure 2. Three ways of selecting diverse yet druglike monomers.

SIMILARITY MEASURES

Molecules are represented by a 1024-bit fingerprint, using the DAYLIGHT fingerprint.[8] Two similarity measures are used. In the case of monomer and SDF whole-molecule comparison, the traditional Tanimoto index expressed by

$$\text{Tanimoto similarity score} = (A \,\&B)/(A + B - A \,\&\, B)$$

is used, where *(A & B)* denotes the number of common bits turned on in both molecules, *A* represents those turned on in molecule *A, and B* denotes those turned on in molecule *B.*

In the case of monomer and SDF substructure comparison, the substructure similarity index

$$\text{Substructure similarity score} = (A \,\&\, B)/A$$

is used, where *A* denotes the number of bits turned on in monomer *A*, and *(A & B)* denotes the number of common bits turned on in both monomer *A* and SDF molecule *B.* The occurrence of the substructure is recorded if this similarity score is greater than 0.9.

This is illustrated in Figure 3, where the boldface 1’s denote those “on” bits that were used in the similarity calculation. The second case clearly shows that the monomer and the whole SDF molecule are not similar if measured by the Tanimoto index; however, the substructure index shows that a part of the SDF molecule is similar to the monomer.

CONSTRUCTION OF DAYLIGHT DATABASES

The process involves collecting from available electronic databases (e.g., Glaxo Wellcome Corporate Database, Spresi[9], etc) and listing all molecules that potentially could be monomers. At this stage availability of the material is not considered; all that is relevant is the fact that someone has recorded its existence in the past. In this manner, five separate DAYLIGHT databases were constructed on the basis of five functional groups: the carboxylic acids (COOH), primary amines (NH_2), secondary amines (NH), aldehydes (CHO), and nitriles (CN). Structures that are unique were subjected to the following analyses:

1. Molecular profiles such as molecular weight, and counts of H-bond donors and acceptors, rotatable bonds, aryl rings, and positive and negative charges are calculated.
2. Jarvis-Patrick cluster analysis is carried out to group structures into clusters on the basis of DAYLIGHT fingerprints.

The calculated properties are then put back into the databases. Thus searches such as diversity selection, similarity search, and molecular profile prescreening can be carried out easily. Once a set has been selected, the detailed information relating to availability can then be searched in the original databases.

Table 1 denotes the number of unique structures considered

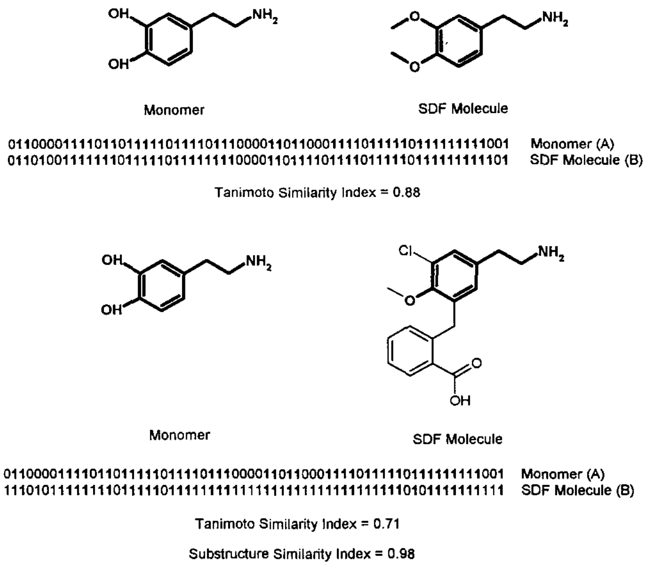

OH
NH2
OH
Monomer
O
O
NH2
SDF Molecule
0110000111101101111101111011100001101100011110111110111111111001 Monomer (A)
01101001111111011111011111111000011011110111101111101111111111101 SDF Molecule (B)
Tanimoto Similarity Index = 0.88
OH
NH2
OH
Monomer
Cl
NH2
O
O
OH
SDF Molecule
0110000111101101111101111011100001101100011110111110111111111001 Monomer (A)
1110101111111011111011111111111111111111111111111110101111111111 SDF Molecule (B)
Tanimoto Similarity Index = 0.71
Substructure Similarity Index = 0.98

Figure 3. Illustration of Tanimoto similarity and substructure similarity indices for calculating whole-molecule and substructural similarities.

and the number of clusters obtained. An illustration of the aldehyde monomers database is shown in Color Plate 1.

RESULTS OF DIVERSE MONOMER SELECTION BASED ON DRUGLIKE MOTIFS

The results are illustrated by an aldehyde monomer set selection. The aldehyde monomers were clustered using the Jarvis-Patrick algorithm under DAYLIGHT with need and near parameter of 10/16 and a fingerprint length of 1024 bits. Some 3200 cluster centroids and 3000 singletons that satisfy a molecular mass of less than 300 Da were selected from the aldehyde database. The following comparisons are made.

Monomer and Standard Derwent File whole-molecule similarity

The 6000 monomers were compared to each of the 30 000 SDF molecules; the highest similarity score for each monomer versus its cluster size is shown in Figure 4.

A general trend was observed, in that as the cluster size decreases, corresponding to an increase in cluster number (Figure 4b), the similarity of the monomer to SDF molecules decreases. Thus, in general, those structurally less common monomers such as singletons from Jarvis-Patrick analysis may not be desired when designing libraries on the basis of known drug motifs.

Table 1. Number of structures contained in the five monomer databases

Database	Unique structures; number of clusters
Monomers_COOH	260 000; 23 000
Monomer_NH	78 000; 7 200
Monomer_NH2	50 000; 4 400
Monomer_CHO	35 000; 3 200
Monomer_CN	115 000; 10 000

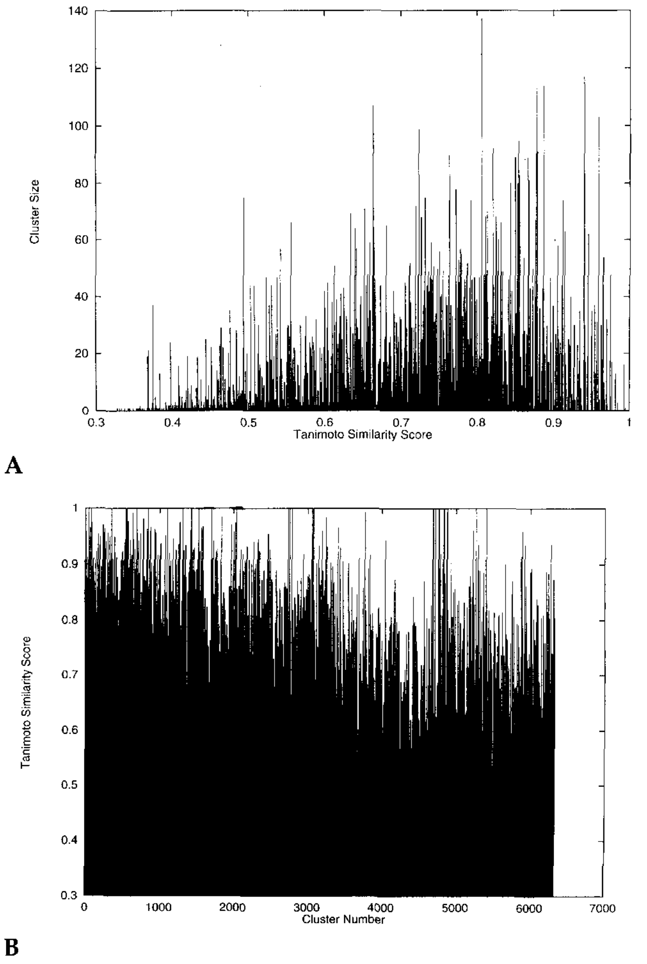


Figure 4. (a) Number of structures in a cluster versus the Tanimoto similarity score between the cluster centroid and its most similar SDF molecule. (b) Tanimoto similarity score between a cluster centroid and its most similar SDF molecule versus the cluster number. Increase in cluster number corresponds to a decrease in cluster size.

Examples of some monomers and their corresponding most similar SDF molecules are shown in Figure 5.

Monomer and exact/similar substructure occurrence in Standard Derwent File molecules

When comparing monomers with substructures of the SDF molecules, similar substructure occurrence is counted if the substructure similarity index is greater than 0.9. The cutoff value is determined by looking at substructures thus selected and is therefore empirical and somewhat arbitrary. Exact substructure occurrence is a specific case of the substructural similarity when the substructural similarity index is 1.0. Exact

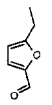
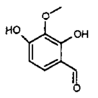
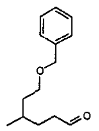
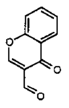
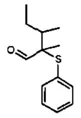
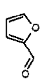
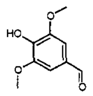
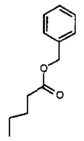
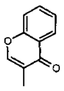
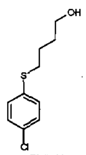
Monomer	Monoomer	Monomer	Monomer	Monomer
				
0.81 FURFURAL	0.94 SYRINGALD	0.88 BENZYLVAL	0.89 MECHROMON	0.66 W2719
				

Figure 5. Examples of monomers and their corresponding most similar SDF molecules. Numbers denote Tanimoto similarity score.

and similar substructural occurrences are not well correlated, as illustrated in Figure 6.

A trend similar to that observed in monomer and whole SDF molecule comparison was also observed here, i.e., in general, as the monomers become singleton-like, the occurrences of similar substructures in SDF become fewer (see Figure 7).

In fact, a general correlation between the whole-molecule similarity and substructural similarity occurrence exists, as one would expect (Figure 8). Thus, using whole-molecule similarity as a measure of closeness to druglike motifs may be a reasonable approximation.

MONOMER SELECTION

Since it appears that whole-molecule and substructural similarity is somewhat correlated, we have selected 630 monomers having a whole-molecule similarity score of greater than 0.8. Approximately half of this set also scores highly as the most frequently occurring substructures in the SDF molecule. Some examples of the potential monomers for library synthesis are shown in Figure 9, along with their whole-molecule similarity score and similar substructure occurrence data.

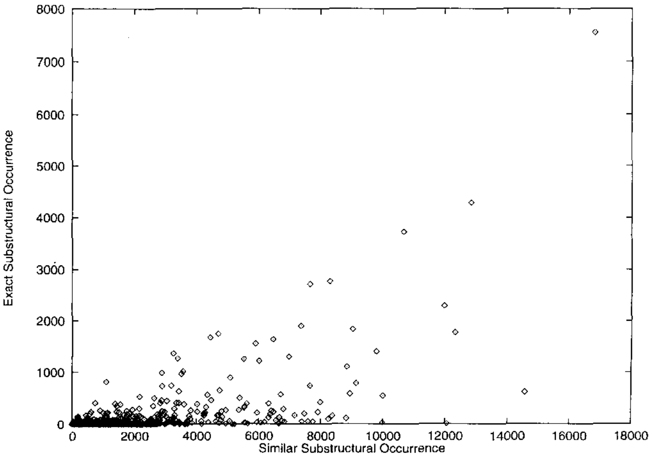


Figure 6. Exact substructure versus similar substructure occurrence of cluster centroid monomers within SDF molecules.

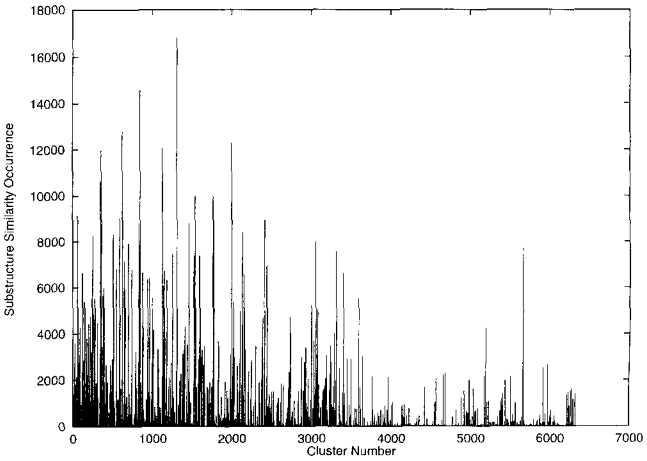


Figure 7. Similar substructure occurrence of the centroid monomers in SDF versus cluster number. Increasing cluster number corresponds to decrease in cluster size. Thus "singleton" monomers occur less frequently.

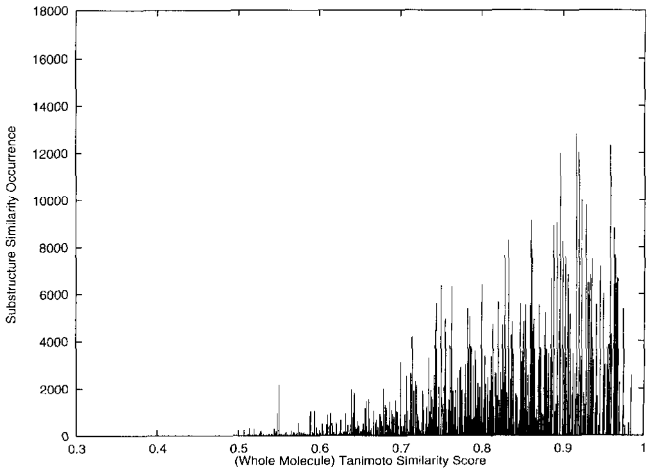


Figure 8. Correlation between similar substructure occurrence and whole-molecule similarity.

CONCLUSIONS

Several lessons have been learned from this exercise.

1. Building central databases of common monomers with predefined clusters and chemical descriptors for similarity/diversity searching and prescreening is extremely valuable in generating ideas and cuts down on preprocessing time. A task that traditionally would have taken a competent computational chemist at least a day to investigate all of the available databases to select, prescreen, and cluster the desired monomers can be done in a matter of seconds, thus increasing business efficiency. Databases like these are being actively used in selecting diverse monomers for primary

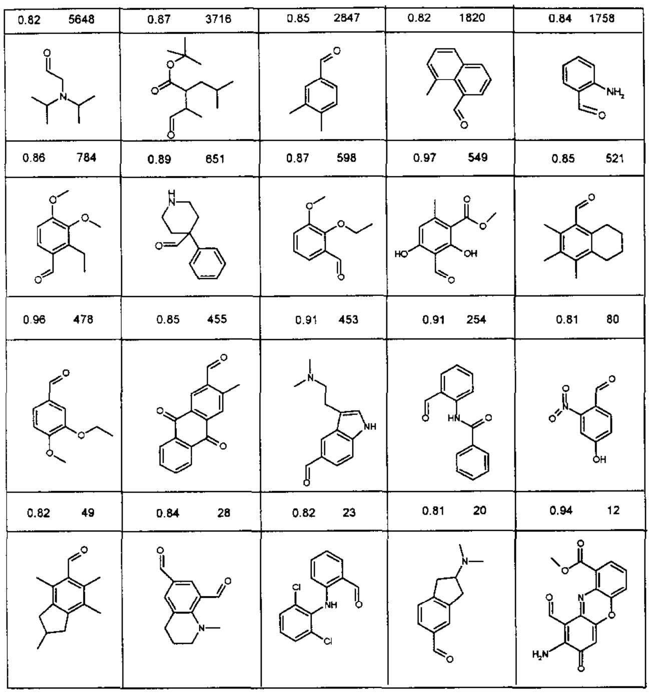


Figure 9. Examples of "druglike" monomers. First number denotes its similarity (in terms of its Tanimoto similarity score) to the most similar SDF molecule. Second number denotes the number of times the monomer occurs as a similar substructure in the SDF molecules.

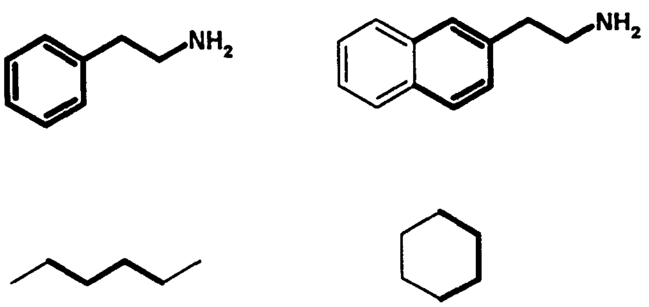
library syntheses, for docking into protein-binding sites for target-oriented projects, and for selecting second-generation library monomers to optimize activities.

2. The idea of selecting structural representatives and comparing them with SDF molecules to reveal similarity is useful in designing diverse yet drug motif-based monomers.
3. In general, commonly occurring structures as represented by the size of a cluster from Jarvis-Patrick cluster analysis tend to be more similar to SDF drug molecules, whereas those structurally less common types such as those represented by singletons tend to be less similar to SDF molecules. Thus two lines of argument can be put forward in selecting monomers:

a. Select those that are structurally common so as to give the libraries a better chance to produce druglike molecules.
b. Select those structurally less common monomers to explore a diversity space that has either previously not shown activity or never been explored. This may or may not result in a fruitful outcome.

We have taken option a for the aldehyde example. However, option b could also be considered in future.

4. In the analysis for substructural similarity, the inherent drawback is the picking up of substructures that may not be desired. Examples are shown here, where naphthalene analogs are identified when we desired only to find the phenyl analogs, or cyclic hexyls are picked while alkyl butyl analogs are desired:



In fact, when the monomer centroids were compared with the SDF molecules for substructure occurrences, those simple chain monomers tend to be counted as frequently occurring owing to the path-matching problem illustrated above.

To overcome this drawback, we are developing a "retrosynthetic fragmentation" method, which should give a more accurate account of the substructure occurrences.

5. We have used the cluster centroids as representatives for monomers. However, by virtue of its definition, the centroids tend to be poor functionality-containing molecules. Thus it may be worthwhile to select those monomers that are some distance away from the centroid as they usually contain more "interesting" functional groups.

In summary, this article describes a methodology developed to select diverse monomers based on drug motifs and a practical application in which aldehyde monomers are selected. It also illustrates the power of setting up central databases containing frequently required monomers with cluster and profile information to increase business efficiency.

ACKNOWLEDGMENTS

The authors thank their colleagues for contributing ideas and comments; particularly Mark Beston, John Bradshaw, Darko Butina, Brian Evans, Darren Green, Mike Hann, and Duncan Judd.

REFERENCES

1 Martin, D.J., Blaney, J.M., Siana, M.A., Spellmeyer, D.C., Wong, A.K., and Moos, W.H. Measuring diversity: Experimental design of combinatorial libraries for drug discovery. *J. Med. Chem.* 1995, **38**(9), 1431–1436

2 Weininger, D., Delany, J., *Daylight Software Manual.* Release 4.51, 1997. Daylight Chemical Information Systems, Inc., 419 East Palace Avenue, Santa Fe, New Mexico 87501, USA

3 Jarvis, R.A. and Patrick, E.A. Clustering using a similarity measure based on shared near neighbors. *IEEE Trans. Comput.* 1973, **C22,** 1025–1034

4 Davies, K. and Briant, C. Combinatorial Chemistry Library Design Using Pharmacophore Diversity, MGMS Meeting, Leeds, 1995

5 Tripos. Molecular diversity manager generates lead followup synthesis candidates. *Tripos Tech. Notes,* 1995, **1**(2)

6 *Cerius 2*. MSI, http://www.msi.com/marketing/products/cchem.html

7 Pearlman, R. Novel software tools for addressing chemical diversity. *Network Sci.* Vol 2, Issue 6/7, June 1996

8 Derwent Information, 14 Great Queen Street, London WC2B 5DF, England, 1993 DAYLIGHT version containing 37 000 structures, supplied by DAYLIGHT.

9 InfoChem GmbH. Spresi 95 is a collection of 3.4 million substances assembled by VINITI of USSR and ZIC of Berlin. Database is available via DAYLIGHT.