Clustering of Large Databases of Compounds: Using the MDL “Keys” as Structural Descriptors

Malcolm J. McGregor* and Peter V. Pallai†

Procept, Inc., 840 Memorial Drive, Cambridge, Massachusetts 02139

Received November 7, 1996⊗

An analysis of chemical structures from several commercially available libraries of compounds is presented with a view of acquiring compounds for screening. The Jarvis−Patrick clustering method has been applied, using the MDL “keys” as structural descriptors. The nature of the MDL keys is examined in this context, some features of the clustering algorithm are discussed, and clustering statistics are presented.

INTRODUCTION

A traditional approach to drug-discovery has been the random screening of large numbers of compounds in a primary assay. Although other methods such as structure-based design[1,2] have made rapid progress, so has the technology for high-throughput screening (HTS) and combinatorial chemistry.[3] HTS creates great demands for compounds to be screened, and to meet this demand there are currently many commercial sources of libraries of compounds. However, the cost of acquiring these can be a large part of the overall research cost for a drug-discovery project. Therefore it is worthwhile to apply the computational resources routinely used in the structure-based approaches to analyze these libraries and make a rational choice about which compounds to purchase from which sources, so as to maximize cost efficiency and number of hits generated.

We have analyzed the chemical structures of several commercial libraries with a view to acquiring compounds for screening. However, it should be noted that a purchasing decision is complicated by many other factors since commercial sources of compounds vary considerably in quality and scope. Some companies supply drug-like compounds specifically for the purpose of drug-discovery, while traditional chemical companies largely supply reagents for chemistry. Some companies pride themselves on careful design of libraries and strict quality control, while with others the source of the compounds can be difficult to establish, the purity may be questionable and, the diversity may not be optimal. There may be overlap between sources, for example, when some of the chemical stock of one company is transferred to another. With some sources one may not be able to choose individual compounds; some compounds may be available only on 96-well plates or in large batches. Last but not least there is the consideration of price, which may be negotiable for a large order. This paper considers some of the issues that can be addressed by looking at the actual structures.

Currently there is considerable interest in the concept of molecular diversity.[3−6] The concept is that to maximize the number of hits in an assay with a target of unknown structure one has to optimize the sampling of “diversity space”. Therefore a reasonable starting point would be

1. define the structures with drug-like or otherwise-desirable properties;
2. optimize the diversity of the set of compounds which satisfy the first criterion;
3. make sure these compounds are sufficiently different from the ones already tested.

This led us to implement a measure of molecular similarity, for which we chose the Tanimoto coefficient using the MDL “keys”[7−9] (see methods section), and a method of clustering, for which we chose the Jarvis−Patrick method.[10] These are commonly used methods, but to our knowledge they have not been extensively applied to large libraries of real compounds. We have analyzed nine databases from commercial sources; for comparison we also included the ACD and CMC databases from MDL (see Methods).

METHODS

We have analyzed nine libraries of compounds which are commercially available. Details are given in Table 1; we wish to thank these companies for the work involved in supplying us with this information. In addition we have included two databases for the purposes of comparison:

1. The Available Chemicals Directory (ACD) compiled by MDL:[7] The ACD is a compilation of commercially available compounds from many different sources, including some of the others considered here. It is the largest database considered here, and it can be regarded as a reference with regard to the kinds of compounds which are commercially available in general.
2. The Comprehensive Medicinal Chemistry database, also from MDL: The CMC contains compounds which have been marketed as drugs. As such these molecules are not necessarily commercially available but they can be used as a reference for molecules with drug-like properties.

We have chosen as a description of molecular structure the MDL keys.[7−9] These were developed for substructure searching but are potentially useful in structure comparison. They are a set of molecular descriptors which are either present or absent in a molecular structure. There are two

* To whom correspondence should be addressed. E-mail: mcgregor@procept.com.

† New address: ArQule, Inc., 200 Boston Avenue, Medford, MA 02155.

⊗ Abstract published in *Advance ACS Abstracts,* May 1, 1997.

S0095-2338(96)00151-5 CCC: $14.00

Table 1. Libraries Used in This Study

name as used in this study	size	source/comments
acd	161171	Available Chemicals Directory from MDL, compilation of commercially available compounds from several sources
bionet	9489	Bionet, commercially available libraries
brandon	50353	Brandon, commercially available libraries
cmc	6683	Comprehensive Medical Chemistry from MDL, database of drug compounds
comgenex	13876	Comgenex, commercially available libraries
maybridge	49802	Maybridge, commercially available libraries
microsource	5198	Microsource, commercially available libraries
mss	7000	Molecular Screening Services, commercially available libraries
optiverse	20018	Optiverse, commercially available combinatorial libraries from Panlabs/Tripos
rrl	14103	Receptor Research Ltd., commercially available libraries
specs	4000	Brandon/SPECS/BioSPECS, commercially available libraries on 96-well plates (sample)

sets of keys: a set of 166 keys which are small topological substructure fragments and a 960 key set which includes the 166 keys but which adds algorithmically generated more abstract atom-pair descriptors. Stereochemistry is not included in either set. For the purposes of comparison each molecule can be regarded as a binary bitstring of 1's and 0's—a "molecular fingerprint"—which facilitates computation. The keys are assigned automatically to every compound in an MDL database, or they can be generated from structure files (e.g., SD files) using the "keyaccs gateway". We applied a program to strip counterions from SD files where necessary. To this end the ACD and CMC were downloaded as SD files, and if the structure consisted of more than one fragment, only the first fragment was used in the structure calculations.

As a measure of similarity between structures we have chosen the binary Tanimoto coefficient, defined by

$$\frac{N_{\mathrm{A\&B}}}{N_{\mathrm{A}} + N_{\mathrm{B}} - N_{\mathrm{A\&B}}}$$

where N_{A} and N_{B} are number of bits set in bitstrings A and B, respectively, and $N_{\mathrm{A\&B}}$ is the number of bits which are common to both. Thus the measure is between 0.0 and 1.0, where 1.0 indicates perfect similarity of the bitstrings. With the MDL keys this usually means that the structures are identical apart from stereochemistry which is not taken into account by the keys.

It is not our intention to discuss the nature of the chemical clustering problem nor the merits of the various clustering methods. The reader is referred elsewhere.[10−15] In general it should be appreciated that chemical structures place extremely tough demands on a clustering algorithm, and there is no objectively correct solution to the chemical clustering problem. Different results will be produced by variations in the algorithm, parameters, similarity measures, or structural descriptors used. It is considered legitimate to choose or adjust these to give the kind of results which are appropriate to the problem. The clustering method chosen was the Jarvis−Patrick method.[10] This has been shown to perform well for chemical clustering;[11−14] it is widely used and is computationally efficient for large databases. One report[15] indicates that some hierarchical clustering methods may perform better at selecting active compounds, but these are usually slower and not so suited to large data sets.

The Jarvis−Patrick algorithm proceeds as follows. The first step is to calculate, for each structure, its list of *J* nearest neighbors. According to the original Jarvis−Patrick definition, two structures cluster together if

1. they are in each others list of nearest neighbors, and
2. they have at least *K* of their *J* nearest neighbors in common, where these are adjustable parameters; we settled on values of 10 and 16, respectively.

Several variations are possible; the only variation we employed is that requirement 1 was relaxed so that only one structure needs to appear in the others list of nearest neighbors.

For each cluster, a loosely termed "centroid" can be determined, from the following

$$\frac{\sum_{j=1,n} (1.0 - T_{ij})^2}{n - 1}$$

where *n* is the number of members in a cluster and T_{ij} are the Tanimoto scores. The structure with the lowest such value is assumed to be the one which is most representative of the cluster.

Generating the nearest neighbor list involves calculating the Tanimoto score for each pair of structures and is therefore an O(N^2) problem, but it only needs to be done once and the results are stored. For the ACD (161 171 structures, 960 keys) this took about 3.5 days on a Silicon Graphics workstation with a 4400 processor. The clustering stage is much faster, so experimenting with different clustering parameters is feasible. Clustering the ACD took about 2.5 h, although the smaller databases take a few minutes. The software for the scoring, clustering, and database comparisons was developed in house; it was written in the C programming language to take advantage of the bitwise operators inherent in the language to maximize speed.

RESULTS AND DISCUSSION

How Well Do the Keys Perform? A Test Case. One of the initial questions was how well the MDL keys and Tanimoto coefficient perform as measures of structural similarity. In particular we wanted to evaluate whether the complete set of 960 keys would give superior results compared to the subset of 166, given that considerable computation time could be saved by using the smaller set. Initial examples suggested that intuitively the 960 keys did give better results, so a way was sought to quantify this. We calculated the Tanimoto coefficient for pairwise comparisons of the 20 natural amino acids. An objective measure of similarity of the amino acids is the frequency of mutation seen in protein sequences, where conservative mutations are more common than mutations between structurally dissimilar

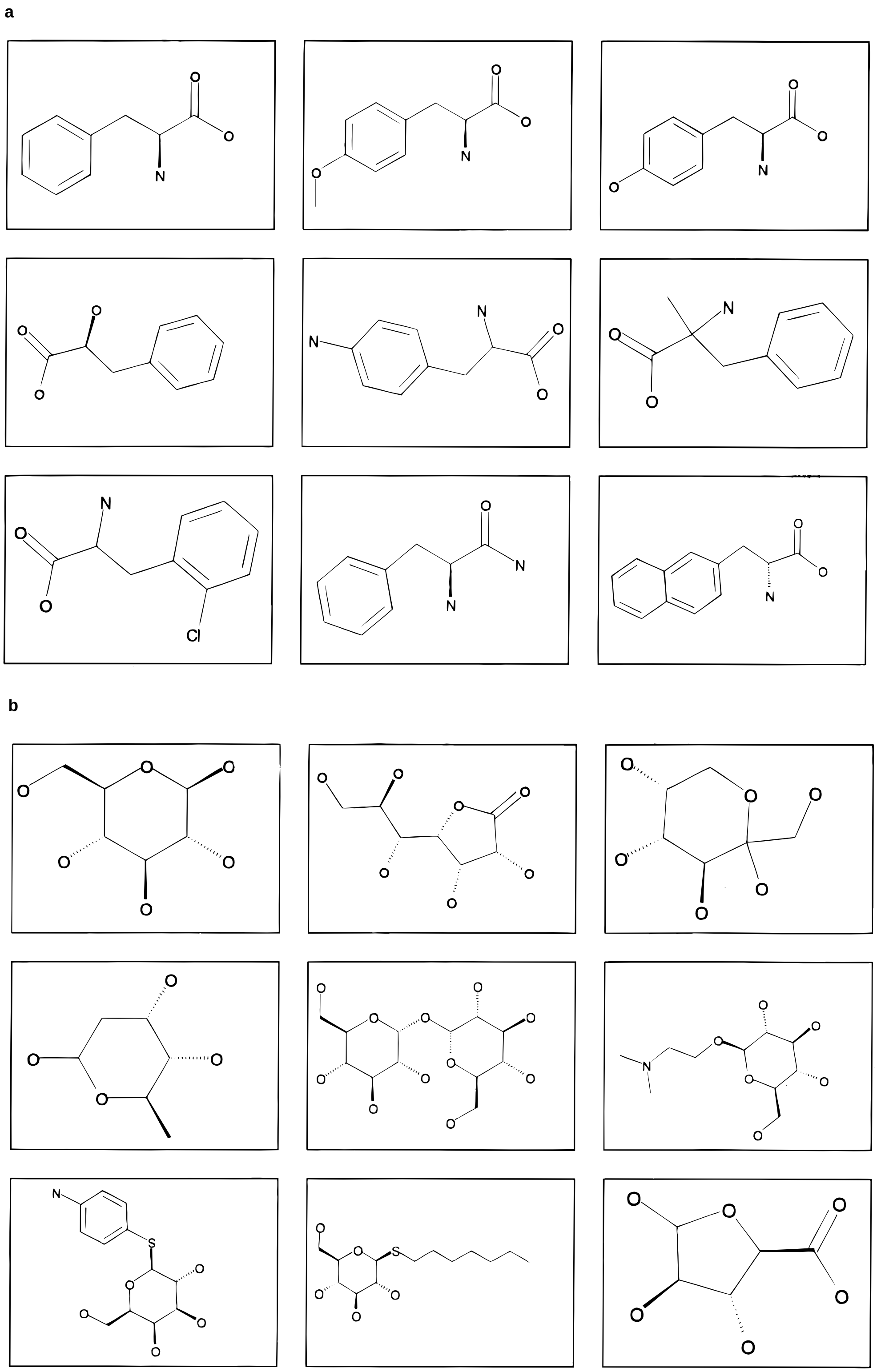

Figure 1. a. Examples of structures from the ACD which have a Tanimoto score of ≥0.75 compared to L-phenylalanine (top left). b. Examples of structures from the ACD which have a Tanimoto score of ≥0.75 compared to β-D-glucose (top left).

residues. This is given by the Dayhoff matrix[16] which is commonly used for protein sequence alignment; we used a more recently calculated version.[17] A correlation coefficient was calculated between the mutation matrix values and the Tanimoto scores for the 190 possible amino acid comparisons using each key set. The results are 0.35 for 166 keys and 0.43 for the 960 key set. The correlations are not high, but it shows that there may be an advantage in using the complete set of keys.

It was considered that the correlations could be improved if certain amino acids were eliminated which have properties in the context of a protein which cannot be predicted from the isolated amino acid structure alone. For example, cysteine forms covalent bonds, and proline and glycine have unique conformational properties in the polypeptide backbone. When cysteine is removed from the calculation the correlations are 0.40 and 0.47 respectively, and when cysteine, proline, and glycine are removed the correlation is

Table 2. Percentage of Molecules with a Druglikeness Index < 10.0 (See Text)

acd	9.8	microsource	3.6
bionet	1.7	mss	1.8
brandon	2.6	optiverse	0.1
cmc	2.9	rrl	0.7
comgenex	3.9	specs	1.3
maybridge	5.0	tripep	0.0

0.46 and 0.53, showing a progressive improvement, with the larger set of keys performing better. With the remaining amino acids the hydrophobic ones which form the core of a protein structure are more strongly conserved than the hydrophilic ones which tend to occur on the solvent-exposed surface. Using the hydrophobic amino acids (alanine, isoleucine, leucine, methionine, phenylalanine, tryptophan, tyrosine, and valine—eight structures, 27 comparisons) the correlations are 0.77 and 0.83. These correlations are now very good, and the difference between them follows the same trend. Using the hydrophilic amino acids (arginine, asparagine, aspartate, glutamine, glutamate, histidine, lysine, serine, and threonine-nine structures, 36 pairs) the correlations are 0.54 and 0.50. This is not as good as the hydrophobic amino acids but better than the complete 20, but now curiously the smaller set of keys perform better. However, it was concluded that in general there may be an advantage in using the complete set of keys; therefore, all our calculations used the complete set since the increased computation time was not prohibitive. If computational resources are limited, the bitstring can be reduced by folding it on itself, although we did not do this.

"Druglikeness". Since choosing diverse or representative molecules only makes sense if those molecules have desirable general properties and since commercial libraries contain a wide variety of compounds sold for different purposes, a way was sought to characterize "druglikeness" using the MDL keys. For this purpose the CMC was used as a reference for drug-like molecules, and the frequency of occurrence of each key was calculated as the fraction of molecules which are positive for that key in the CMC database. As a reference for nondruglike molecules a version of the ACD was prepared where molecules were eliminated which have a Tanimoto score of ≥0.75 compared to any molecule in the CMC database. This eliminated 24% of the ACD as in Table 4. A reasonable measure of druglikeness was postulated to be

$$\sum_{n=1,\,960} \text{keys}[n]*(\text{CMC}[n] - \text{ACD}[n])$$

The measure is a sum over the 960 keys, where keys[] is the bitstring of 1's and 0's for a molecule, CMC[] is the frequency of occurrence of the keys in the CMC database, and ACD[] is the frequency of occurrence of the keys in the "nondruglike" subset of the ACD as described above. In general this produces scores in the range −10 to +60. An arbitrary cutoff of <10 was used to identify molecules at the low end of the range. These are seen to be generally small, uninteresting molecules or ones with unusual atoms or groups, such as metals. The upper end of the scale (above, say 50) not only contained larger, more highly functionalized structures which are generally more desirable, but it also included peptides and high molecular weight natural products; there are less desirable but could be screened out with a molecular weight cutoff. Table 2 shows the percentage of molecules for each database with a score of <10. In general they are in the low single figure range, but for the

Table 3. Overlap between Databases: Tanimoto = 1.0[a]

	n	acd	bio	bra	cmc	com	may	mic	mss	opt	rrl	spe
acd	161171		5	3	2	2	29	1	0	0	7	0
bionet	9489	79		1	0	0	2	0	0	0	5	0
brandon	50353	9	0		0	2	3	0	0	0	1	5
cmc	6683	29	0	2		1	2	7	0	0	0	1
comgenex	13876	16	0	5	0		3	2	0	0	0	1
maybridge	49802	92	0	3	0	1		0	0	0	21	0
microsource	5198	24	0	3	9	6	4		0	0	0	0
mss	7000	3	0	1	0	1	1	0		0	0	1
optiverse	20018	0	0	0	0	0	0	0	0		0	0
rrl	14103	79	3	2	0	0	73	0	0	0		0
specs	4000	12	0	55	1	2	2	0	1	0	1	

[a] The numbers are the percentage of structures in the row database which have a match with a structure in the column database (thus the matrix is not symmetric). This involves doing a pairwise comparison between all the entries in one database and all the entries in the other. A match is defined as a Tanimoto score of 1.0, i.e., identical bitstrings (keys).

Table 4. Overlap between Databases: Tanimoto ≥ 0.75[a]

	n	acd	bio	bra	cmc	com	may	mic	mss	opt	rrl	spe
acd	161171		18	46	24	34	63	22	20	16	33	17
bionet	9489	96		48	18	30	69	18	21	26	65	26
brandon	50353	78	33		34	50	62	35	36	30	46	47
cmc	6683	76	12	48		29	41	40	17	20	22	19
comgenex	13876	78	18	56	24		55	32	36	23	34	29
maybridge	49802	99	30	50	17	35		19	21	24	64	23
microsource	5198	81	11	58	41	42	58		21	11	26	19
mss	7000	76	21	56	25	64	56	30		25	37	31
optiverse	20018	54	18	36	22	30	40	14	19		30	20
rrl	14103	97	39	53	21	36	95	20	25	30		28
specs	4000	75	26	85	31	47	52	28	28	23	41	

[a] Data presented as Table 3, but a match is defined as Tanimoto score of ≥0.75, i.e. a certain degree of similarity (see Figure 1).

whole ACD it is 9.8% which indicates that it may be worth screening out these structures.

Databases: Overlap and Overall Similarity. An important initial consideration with the libraries is how much overlap or redundancy there is between them. This is shown in Tables 3 and 4 where the databases have been compared by calculating the Tanimoto score for all pairwise comparisons of structures between two libraries. The entry is the percentage of compounds in the row database which matched a compound in the column database. The matrix is thus assymmetric. The comparisons have been made at two levels. Table 3 gives the overlap at the 1.0 level of Tanimoto score, i.e., only compounds with identical bitstrings are regarded as matches. Note that identical bitstrings do not imply identical structures, since some features such as chirality are not captured by the fingerprint assignment. However, in practice, we note that the MDL keys capture almost all of the structural information for the purposes of comparison. We can then refer to these hits as near-identities, and this provides a good measure of the overlap between databases in terms of common compounds. It can be seen, for example, that 92% of compounds in the Maybridge database are also found in the ACD, which is high compared to the other databases compared to the ACD. This is not surprising since we know that the ACD contains many Maybridge compounds. But in general the overlaps are quite low.

Table 4 gives the overlap between databases compared at the 0.75 level of Tanimoto score. At this level two compounds can be considered to be analogs of each other. Figure 1 gives some examples of molecules which are similar at the 0.75 level; they are generally structures with the same overall framework but may differ in one or two functional groups. This provides a measure of the overlap between databases in terms of the classes of compounds contained. Now there is high overlap between most of the libraries and the ACD, generally 80−90%. This shows that there are not many classes of compounds which are not already represented in the ACD. The lowest overlap with the ACD is with the Optiverse library (54%). This consists of combinatorial libraries of compounds which have been designed to be optimally diverse and are sold specifically for the purposes of screening in a drug-discovery assay. This suggests that it may be a good source of novel compounds if the ACD has already been extensively explored.

Databases: Clustering. Clustering is a commonly used procedure for choosing representative compounds from a set, and the Jarvis−Patrick method described above was applied to the data as follows. Table 5 gives some clustering statistics for each of the databases. In general the number of clusters (of size greater than 1) expressed as a percentage of database size is in a narrow range around 8-10%. This is over a large range of database size. This is probably a feature of the self-scaling nature of the Jarvis−Patrick algorithm. The last column in Table 5 gives the number of singletons as a percentage of database size. Here the results are more variable, generally between 10−30%. Singletons can be regarded either as compounds which are not clustered or as clusters with only one member. At first the number of singletons may appear to be disturbingly high. There is also a generally large number of small clusters and a small number of large clusters. It is difficult to find parameters that result in a more even distribution of clusters. The question arises whether this is an accurate representation of the distribution of molecules in these libraries or if it is an artifact of the clustering algorithm.

Table 5. Clustering Statistics for Databases Clustered Individually[a]

	db size	clusters		singletons	
	n	*n*	%	*n*	%
bionet	9489	746	7.9	1002	10.6
brandon	50353	4381	8.7	7973	15.8
cmc	6683	580	8.7	2128	31.8
comgenex	13876	1131	8.2	3517	25.3
maybridge	49802	4902	9.8	13081	26.3
microsource	5198	480	9.2	1344	25.9
mss	7000	589	8.4	1158	16.5
optiverse	20018	1812	9.1	4372	21.8
rrl	14103	1476	10.5	5174	36.7
specs	4000	348	8.7	934	23.4
db 10	180522	19351	10.7	35272	19.5
acd	161171	15787	9.8	37835	23.5
tripep	8000	760	9.5	119	1.5

[a] The columns are numbers of entries in each database; clusters obtained using methodology described in the text, expressed as an absolute number and as a percentage of database size; singletons expressed as an absolute number and as a percentage of database size. Db10 is the first 10 databases combined and then clustered (forms the basis for the data in table 6). Tripep is the database of 8000 tripeptides consisting of the 20 naturally occurring amino acids.

To address this question a database was prepared of the 8000 possible tripeptides containing the 20 natural amino acids. Here the content of the library is known from the start, and it would be expected that the molecules are likely to be evenly spread throughout a property space. Clustering of this database produced only 119 singletons (1.5% of structures) of which nine were peptides with three amino acids of the same type. This indicates that the large number of singletons obtained with most libraries is not a necessary feature of the clustering algorithm and that a more even distribution can be obtained with an appropriate library. Another feature of this result is that the Tanimoto scoring did not differentiate the same amino acids differently ordered in the sequence. The exception was peptides which contain proline at the N-terminus where the molecules are unique in containing a secondary amine, and these tended to be clustered together. This reflects the way that the keys detect only small substructures and not large scale connectivity.

Table 6 gives the clustering statistics for the nine commercial databases plus the CMC combined and then clustered. The structures which resulted as centroids and singletons have been broken down according to which source they came from. With some sources, for example, the library from Receptor Research Ltd., there is a high ratio of centroids to singletons (8.0:3.3), whereas with others, for example, the CMC, the ratio is low (2.7:7.1). The origin of this difference must reflect the distribution of structures in the libraries, but further work will be needed to better characterize those distributions.

CONCLUSION

To review the three criteria given in the Introduction see the following:

1. *Define the structures with druglike or otherwise desirable properties:* Druglikeness is an elusive concept; the MDL keys provide one way to at least eliminate compounds which are least likely to satisfy this criterion, as described

Table 6. Clustering Statistics for the 10 Shown Databases Combined and Then Clustered[a]

	db size		centroids		singletons	
	n	%	*n*	%	*n*	%
bionet	9489	5.3	930	4.8	1399	4.0
brandon	50353	27.9	4968	25.7	7948	22.5
cmc	6683	3.7	514	2.7	2495	7.1
comgenex	13876	7.7	1257	6.5	3508	9.9
maybridge	49802	27.6	6769	35.0	10401	29.5
microsource	5198	2.9	634	3.3	1285	3.6
mss	7000	3.9	693	3.6	1417	4.0
optiverse	20018	11.1	1821	9.4	5073	14.4
rrl	14103	7.8	1557	8.0	1172	3.3
specs	4000	2.2	208	1.1	574	1.6
total	180522	100.0	19351	100.0	35272	100.0

[a] The columns are database size as a number and as the percentage of the total; centroids expressed as the number of centroids which belong to that database and the percentage of the total number of centroids; singletons expressed as the number of singletons which belong to that database and the percentage of the total number of singletons.

above. Explicit criteria, such as molecular weight or the presence or absence of substructures or functional groups, may also be used to eliminate undesirable structures or to highlight favorable ones.

2. *Optimize the diversity of the set of compounds which satisfy the first criterion:* The results obtained from the Jarvis–Patrick clustering method suggest that choosing the centroids and singletons produces a satisfying set of compounds for testing.

The concept of molecular diversity is presently a topic of much discussion. Probably the most important point is that diversity is not a property in itself but a measure of how properties differ in a set of objects. Thus diversity is context dependent and relates to the properties chosen for measurement. The molecular fingerprinting method used here gives only a distance between points and does not define a property space as such. Thus it is difficult to give a measure of how diverse one library is compared to another, or which regions of property space are occupied by the different libraries. The Jarvis–Patrick method has the advantage that the cluster centroids are representative of the library contents.

We have used a two-dimensional description of chemical structure. It could be argued that for a thorough characterization of molecular properties a 3-D description is needed. However, calculations of 3-D properties of molecules, especially ones involving conformational search, can be prohibitively expensive for large databases. For our purposes, at least for a general statistical treatment, a 2-D description is probably sufficient.

3. *Make sure these compounds are sufficiently different from the ones already tested:* Although we have not addressed this question in the present paper, the Tanimoto scoring and the way we have set up the data provide a straightforward way to do compound by compound comparisons to eliminate similar ones; the results will obviously be context dependent.

ACKNOWLEDGMENT

We would like to thank Paul Kowalczyk and Bert Thomas at Procept and Bill Balke, John McCarthy and Tom Moock at MDL for helpful discussions.

REFERENCES AND NOTES

(1) Verlinde, C. L. M. J.; Hol, W. G. J. Structure-Based Drug Design: Progress, Results and Challenges. *Structure* **1994**, *2*, 577–587.

(2) Whittle, P. J.; Blundell, T. L. Protein Structure-Based Drug Design. *Annu. Rev. Biophys. Biomol. Struct.* **1994**, *23*, 349–375.

(3) Lebl, M.; Leblova, Z.; Dynamic database of references in molecular diversity. Internet http://vesta.pd.com.

(4) 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.

(5) Bemis, G. W.; Murcko, M. A.; The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.

(6) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E.; Neighborhood Behavior: A Useful Concept for Validation of "Molecular Diversity" Descriptors. *J. Med. Chem.* **1996**, *39*, 3049–4059.

(7) MDL Information Systems, Inc., 14600 Catalina Street, San Leandro, CA 94577.

(8) Molecule Database Administration Guide, Ver. 2.0.1, MDL Information Systems, Inc., 1996, pp 2–13, 8–57.

(9) Ahrens, E. K. F. Customization for Chemical Database Applications. In *Chemical Structures*; Warr, W. A., Ed.; Springer: Berlin, 1988; pp 97–111.

(10) Jarvis, E. A.; Patrick, E. A. Clustering Using a Similarity Measure Based on Shared Nearest Neighbors. *IEEE Trans. Comput.* **1973**, C-22, 1025–1034.

(11) Barnard, J. M.; Downs, G. M. Clustering of Chemical Structures on the Basis of Two-dimensional Similarity Measures. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 644–649.

(12) Willet, P. *Similarity and Clustering in Chemical Information Systems*; Research Studies Press, Wiley: New York, 1987.

(13) Willett, P.; Winterman, V.; Bawden, D. Implementation of Nonhierarchical Cluster Analysis Methods in Chemical Information Systems: Selection of Compounds for Biological Testing and Clustering of Substructure Search Output. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 109–118.

(14) Willett, P.; Winterman, V. A Comparison of Some Measures for the Determination of Inter-Molecular Structural Similarity. *Quant. Struc.-Act. Relat.* **1986**, *5*, 18–25.

(15) 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.

(16) Dayhoff, M. O.; Schwartz, R. M.; Orcutt, B. C. *In Atlas of Protein Sequence and Structure. National Biomedical Research Foundation, Washington, DC.* 1978; Vol. 5, Suppl. 3, pp 345–352.

(17) Jones, D. T.; Taylor, W. R.; Thornton, J. M. The Rapid Generation of Mutation Data Matrices from Protein Sequences. *Comput. Appl. Biosci.* **1992**, *8*, 275–282.

CI960151E