Rapid Quantification of Molecular Diversity for Selective Database Acquisition

David B. Turner, Simon M. Tyrrell, and Peter Willett*

Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield, Western Bank, Sheffield S10 2TN, United Kingdom

Received May 25, 1996⊗

There is an increasing need to expand the structural diversity of the molecules investigated in lead-discovery programs. One way in which this can be achieved is by acquiring external datasets that will enhance an existing database. This paper describes a rapid procedure for the selection of external datasets using a measure of structural diversity that is calculated from sums of pairwise intermolecular structural similarities.

INTRODUCTION

Databases of chemical structures play an important role in the development of novel pharmaceuticals and agrochemicals.[1] Thus far, organizations have paid the most attention to the molecules contained within their own corporate databases; however, these will typically contain only a limited number of structural types and there is thus much interest in techniques that can augment a corporate database by increasing the diversity of the molecules that are available for biological testing. Additional molecules can come from a range of sources, including publicly available commercial databases, collaborations with academic synthetic groups, combinatorial chemistry and *de novo* design programs, folk medicine, specialist synthetic organizations, and compound-exchange agreements with other organizations. In what follows, we shall refer to any set of structures that are possible additions to a company's existing corporate database as an *external dataset*.

The increasing range of sources has resulted in a need for quantitative measures of the diversity of an external dataset and of the extent to which the acquisition of that dataset will increase the diversity of an existing database. There have already been several descriptions of measures of structural diversity;[2−6] this communication reports a further such measure and an algorithm for its calculation that allows it to be applied to even the largest datasets at minimal computational cost.

MEASUREMENT OF STRUCTURAL DIVERSITY

The Measure. Martin *et al.*[2] have described several ways of estimating structural diversity in the design of peptoid-based combinatorial libraries. One of their approaches involves a diversity measure based on a matrix that contains all of the pairwise structural similarities for a set of molecules, and we have adopted this idea in the work reported here. Specifically, we suggest that the diversity, *D(A)*, of a database, *A*, containing *N(A)* molecules, should be defined to be the mean pairwise intermolecular dissimilarity, *i.e.*,

$$D(A) = 1 - \frac{\sum_{J=1}^{N(A)}\sum_{K=1}^{N(A)} \mathrm{SIM}(J,K)}{N(A)^2}$$

where SIM(*J*,*K*) is the similarity between two molecules, *J* and *K*, in *A* calculated using some measure of intermolecular structural similarity. If an appropriately normalized similarity coefficient is used for the calculation of the SIM(*J*,*K*) values then

$$0 \leq D\,(\mathrm{A}) \leq 1$$

A value of 1 for *D*(*A*) corresponds to a database in which all of the molecules have a zero-valued similarity with each other, *i.e.*, a database that is as diverse as possible given the structural attributes that have been used to characterize each of the molecules, while a value of 0 for *D(A)* corresponds to a database in which all of the molecules have identical descriptions. In practice, of course, the range of feasible values for *D*(*A*) is very much less than unity.

Similarity measures based on fragment bit-strings are widely used for similarity searching, clustering, and dissimilarity-based compound selection in databases of 2D structures[7,8] and analogous, distance-based bit-string representations have been suggested for similarity searching in databases of 3-D structures.[9] In both cases, the similarity between a pair of molecules is derived from the number of bits common to the bit-strings representing two molecules, using one of a range of available similarity coefficients,[10] and we have used this approach for all of the experiments reported here. However, the diversity measure is applicable to any situation in which a molecule is represented in vector form, *e.g.*, by sets of topological indices[11,12] or calculated molecular properties.[13,14]

The Algorithm. The need to calculate all of the pairwise similarities means that the calculation of *D*(*A*) has an expected time complexity of O($N(A)^2$), and this will clearly cause substantial problems with databases containing tens, or hundreds, of thousands of molecules. However, there is an alternative, O(*N*(*A*)) algorithm that can be used if the cosine coefficient[10] is used for the calculation of the individual SIM(*J*,*K*) values in the numerator of the diversity measure.

* To whom all correspondence should be addressed. E-mail P.WILLETT@SHEFFIELD.AC.UK.

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

S0095-2338(96)00463-5 CCC: $14.00

Let each molecule, J, in the database A be represented by a vector in which the Ith element, $M(J,I)$, represents the weight of the Ith feature in $M(J)$. Let A_C be the linear combination, or *centroid*, of the individual molecule vectors $M(J)$ ($1 \le J \le N(A)$), with $W(J)$ being the weight of the Jth vector in A, so that the Ith element of A_C, $A_C(I)$, is given by

$$A_C(I) = \sum_{J=1}^{N(A)} W(J) \times M(J,I)$$

The dot product of the vector A_C with itself is given by

$$\text{DOTPROD}(A_C, A_C) = \left(\sum_{J=1}^{N(A)} W(J) \times M(J)\right) \cdot \left(\sum_{K=1}^{N(A)} W(K) \times M(K)\right)$$

which may be rewritten as

$$\text{DOTPROD}(A_C, A_C) = \sum_{J=1}^{N(A)} \sum_{K=1}^{N(A)} W(J) \times W(K) \times M(J) \cdot M(K)$$

Now

$$M(J) \cdot M(K) = \sum_{I=1}^{F} M(J, I) \times M(K, I)$$

where F is the number of features in the vector representing each molecule, and thus

$$\text{DOTPROD}(A_C, A_C) = \sum_{J=1}^{N(A)} \sum_{K=1}^{N(A)} \sum_{I=1}^{F} W(J) \times W(K) \times M(J, I) \times M(K, I)$$

Following earlier work by Voorhees,[15] Holliday *et al.*[16] suggested that the weights $W(J)$ should be

$$W(J) = 1 \Big/ \sqrt{\sum_{I=1}^{F} M(J, I)^2}$$

i.e., the reciprocal square root of the squared elements of the vector $M(J)$ and similarly for $W(K)$. Hence,

$$\text{DOTPROD}(A_C, A_C) = \sum_{J=1}^{N(A)} \sum_{K=1}^{N(A)} \left[\sum_{I=1}^{F} M(J, I) \times M(K, I) \Big/ \sqrt{\sum_{I=1}^{F} M(J, I)^2 \times \sum_{I=1}^{F} M(K, I)^2} \right]$$

The bracketed function on the right-hand side of the expression above is simply the cosine coefficient, $\cos(J,K)$, for the similarity between the molecules J and K and thus

$$\text{DOTPROD}(A_C, A_C) = \sum_{J=1}^{N(A)} \sum_{K=1}^{N(A)} \cos(J, K)$$

The sum of all of the pairwise cosine similarities for the molecules in A is hence given by the dot product of the vector centroid of A with itself if, and only if, the individual molecule vectors are weighted using the reciprocal square-root weighting scheme. The diversity, $D(A)$, is then obtained by dividing this dot product by $N(A)^2$ to give the mean similarity when averaged over all pairs of the molecules in A and subtracting the result from one, *i.e.*,

$$D(A) = 1 - \frac{\text{DOTPROD}(A_C, A_C)}{N(A)^2}$$

The use of this particular weighting scheme hence results in a linear algorithm, *i.e.*, it has an expected running time proportional to $N(A)$, since each of the molecules in A must be processed to calculate the centroid A_C. This is in marked contrast to the quadratic algorithms that are generally required for the calculation of intermolecular similarities, *e.g.*, Shemetulskis *et al.*[3] report run-times of tens of CPU days (on analogous equipment to that used here) for their cluster-based approach to enhancing the diversity of the corporate database at Parke-Davis.

SELECTION OF AN EXTERNAL DATASET

The analysis above provides a very fast way of calculating all of the pairwise similarities between the molecules within a given database and hence of calculating the diversity of that database. The same approach can also be used to quantify the change in diversity that occurs when an external dataset X, containing $N(X)$ molecules and with a centroid X_C, is added to an existing database, A, to yield a new, merged database, AX containing $N(A) + N(X)$ molecules and with a centroid AX_C. The diversities of A, X, and AX are

$$1 - \frac{\text{DOTPROD}(A_C, A_C)}{N(A)^2},$$

$$1 - \frac{\text{DOTPROD}(X_C, X_C)}{N(X)^2},$$

and

$$1 - \frac{\text{DOTPROD}(AX_C, AX_C)}{(N(A) + N(X))^2},$$

respectively. The change in diversity of A as a result of adding X, $\delta(A)$, is hence

$$\delta(A) = D(AX) - D(A) = \frac{\text{DOTPROD}(A_C, A_C)}{N(A)^2} - \frac{\text{DOTPROD}(AX_C, AX_C)}{(N(A) + N(X))^2}$$

Now

$$\text{DOTPROD}(AX_C, AX_C) = \text{DOTPROD}(A_C, A_C) + \text{DOTPROD}(X_C, X_C) + 2 \times \text{DOTPROD}(A_C, A_C)$$

Substituting this value for $\text{DOTPROD}(AX_C, AX_C)$ into the equation above for $\delta(A)$, it is possible to calculate the change in diversity that will take place given just a knowledge of the centroids of, and the numbers of molecules in, the external dataset and the original database. This assumes that A and X are disjoint, *i.e.*, that they do not have any molecules

Table 1. Summary Characteristics of the Databases Studied[a]

number of	CAS	COMB	Maybridge	NCI	Starlist	WDI
heavy atoms	19.8 (6.9)	53.7 (30.2)	20.8 (6.1)	18.8 (8.5)	15.0 (5.7)	28.6 (19.4)
rotatable bonds	6.8 (4.3)	25.2 (17.8)	6.0 (3.1)	6.6 (5.0)	5.3 (3.9)	11.6 (11.7)
rings	2.1 (1.4)	4.6 (2.8)	2.3 (1.0)	2.1 (1.5)	1.5 (1.0)	2.9 (2.0)

[a] The figures quoted (to one decimal place) are means and standard deviations (in brackets) when averaged over all of the molecules in a database.

in common, since this would result in the size of the merged database being less than $N(A) + N(X)$. This is, however, not a problem since duplicate molecules can be identified extremely rapidly by dictionary look-up using, *e.g.*, the discriminating index described recently by Hu and Xu.[17]

If several external datasets, X_1, X_2, *etc*. are available, the calculation above can be carried out for each such dataset, thus enabling the identification of that which will best serve to increase the structural diversity of the existing corporate database. The fact that it is not necessary to have access to the individual compounds within each external dataset, X, might be of benefit in cases where the provider of an external dataset wished to minimise the disclosure of structural information prior to the incorporation of that dataset in an existing database.

EXPERIMENTAL DETAILS

The use of the diversity measure is illustrated by calculating the diversities of six databases (five of them publicly available and one of them a combinatorial library) and the changes in diversity that occur when pairs of these databases are merged. The public databases were as follows: a 31 291-molecule random subset of the Chemical Abstracts Service (CAS) database; the 47 165 molecules that comprise the Maybridge database; 117 656 molecules from a very wide range of sources that have been tested for antitumor activity by the National Cancer Institute (NCI); 8152 molecules in the Starlist database for which experimental octanol/water partition coefficients are available; and the World Drugs Index (WDI), which contains 26 866 molecules that have been tested in clinical trials or are currently available as drugs. In addition, a combinatorial library of peptides was built from sets of 400 primary amines, each containing a single primary amine group, and 400 carboxylic acids, each containing a single carboxylic acid group, selected from WDI; all possible pairs of acids and amines were then combined to give a 160 000-molecule library, called COMB.

Some of the principal structural characteristics of the six databases are summarized in Table 1, which lists the means and standard deviations (in brackets) for the following parameters when averaged over all of the molecules for each database: the number of heavy atoms; the number of rotatable bonds; and the number of rings. It will be seen that COMB is very different in character from all of the public databases, with the largest values for all three of the parameters listed. It is followed by WDI, then CAS, Maybridge, and NCI, all of which are very similar in character, and finally by Starlist, which has the smallest values for all three of the parameters.

A database was loaded into the UNITY chemical information management system. Each molecule was represented by the default 2D bit-string fingerprint, which contains 988 bits. The various diversity values were calculated using programs written in C and run under Unix. The longest run, for the calculation of all of the diversity values involving the COMB database, took 15 min of elapsed time on a multiuser Silicon Graphics R4000 workstation, thus demonstrating the great efficiency of the procedures suggested here.

Table 2. Diversity of Single and Merged Databases[a]

	CAS	COMB	Maybridge	NCI	Starlist	WDI
CAS	0.690	−0.052	−0.020	0.010	0.006	−0.023
COMB	0.179	0.460	0.124	0.113	0.228	0.031
Maybridge	0.023	−0.064	0.647	0.045	0.013	0.010
NCI	−0.002	−0.129	−0.010	0.702	0.001	−0.010
Starlist	−0.011	−0.018	−0.047	−0.004	0.706	−0.052
WDI	0.043	−0.133	0.033	0.068	0.030	0.624

[a] The ijth element of the table contains either the diversity of the ith database (when $i = j$) or the change in diversity when the database in the jth column is added to the database in the ith row (when $i \neq j$).

RESULTS AND DISCUSSION

The main experimental results are shown in Table 2. The diagonal elements in this table give the diversity measures for the individual databases, while the off-diagonal elements ij give the change in diversity resulting from the addition of the external dataset in the jth column to the database in the ith row.

Inspection of the diagonal elements suggests that the databases can be ranked in order of decreasing diversity as follows:

Starlist > NCI > CAS > Maybridge > WDI > COMB

It will be seen that the diversity of the combinatorial library, COMB, is far less than for the public databases. This finding is intuitively reasonable in that all of the molecules in COMB share at least some common structural features, with a consequent upperbound on the degree of diversity that is obtainable (and we would expect that this upperbound would be still lower than that observed here if one were to use a combinatorial library based on a large common ring system, such as a benzodiazepine or xanthane library). It is also in agreement with previous studies of public databases and combinatorial libraries, involving alternative measures of structural diversity.[2,4,6] After COMB, the next least diverse databases are WDI and Maybridge, both of which contain sets of compounds of a particular type—drugs (putative or actual) in the case of WDI and samples and intermediates for pharmaceutical research in the case of Maybridge—which would again tend to limit their diversities. Both CAS and NCI would be expected to contain a wide range of structural types given the wide range of sources from which these databases are drawn, and similar comments apply to the Starlist database, where very substantial efforts have been made over the years to measure partition coefficients for as wide a range of compounds as possible.

An inspection of the off-diagonal elements of Table 2 shows that both positive and negative changes in diversity can occur when two databases are merged. For example, a positive change occurs when NCI is considered as an external dataset for addition to the Maybridge database, *i.e.*, there is a reduction in the mean intermolecular similarity and hence an increase in diversity. In fact, for the Maybridge database, NCI would appear to be the most advantageous acquisition. COMB, conversely, causes a marked reduction in diversity when it is merged with Maybridge; indeed, addition of COMB reduces the diversity of all of the public databases considered here, whereas the diversity of COMB is increased substantially if any of the other databases are added to it.

Several alternative approaches to estimating the diversity of a database have been described.[2−6] Of these, the most rapid to compute is probably that advocated by Martin *et al*.,[2] who have suggested counting the number of bits that are set in the union of all of the fingerprints for a database. It would be possible to measure the change in diversity resulting from the merging of two databases simply by comparing the union bit-strings of the original and the merged databases; however, we believe that the approach suggested here is superior for two reasons. Firstly, the number of bits indicates merely that some particular bit is set for at least one molecule within a database but provides no indication as to how frequently this occurs, whereas such frequency information is an inherent part of the measure suggested here. Secondly, a union bit-string provides information only about the individual molecules comprising the database, not about the similarity relationships that exist between pairs of these molecules, whereas this information, in the form of the sums of similarities, is provided here. There is also the practical problem that a database may be sufficiently diverse for all of the bits in the union bit-string to be set, meaning that it is not possible further to increase the diversity irrespective of the content of any additional dataset that is merged with it. This problem was found to occur here with the NCI database.

The main focus of the paper has been the evaluation of entire external datasets but our procedures can also be used to select individual compounds from such a dataset. Assume that x_I is the Ith molecule ($1 \leq I \leq N(X)$) in the external dataset X. The effect of adding each such molecule to A is determined by evaluating the expression for $\delta(A)$ with X being replaced by x_I, *i.e.*, the external dataset is considered to consist of $N(X)$ subdatasets, each containing a single molecule. Those n molecules (where n is a user-defined parameter) in X are then added to A that result in the largest change in the diversity of the latter. In such a case, the formula for $\delta(A)$ can be drastically simplified since both DOTPROD(X_C, X_C) and $N(X)$ will be one and since both DOTPROD(A_C, A_C) and $N(A)$ will be constants. The molecules in X thus need be evaluated solely on the basis of DOTPROD(A_C, X_C), *i.e.*, the similarity between x_I, appropriately weighted, and the centroid of A. This approach is simple in concept; however, it considers only the similarity relationships between A and each individual molecule x_I and ignores the effects on the overall diversity of the pairwise similarities between the molecules that have been extracted from X for inclusion in AX. This can be overcome by using the dissimilarity selection algorithm described by Holliday *et al*.,[16] which tries to identify the most diverse n molecules from a dataset of N molecules and which also uses the centroid approach for the rapid calculation of sums of intermolecular similarities. In the present context, the algorithm would seek to identify the most diverse set of $N(A) + n$ molecules from the $N(A) + N(X)$ molecules in AX, subject to the first $N(A)$ members of this chosen subset being the original members of A.

CONCLUSIONS

This paper has described a simple measure that can be used to quantify the change in structural diversity that will occur when a selective compound-acquisition program is used to increase the range of structural types present in an existing database. The measure provides a single-number description of the similarity relationships that exist within a dataset, and is based on the measures of inter-molecular structural similarity that have been used in previous studies of similarity searching and clustering.

The measure's focus on the characteristics of an entire database means that it is far less discriminating than measures of diversity that characterize individual molecules, such as the HookSpace Index of Boyd *et al*.,[4] the three-point pharmacophore approach described by Ashton *et al*.[5] and by Martin *et al*.,[6] or the modification of the dissimilarity-based selection algorithm of Holliday *et al*. that has been discussed in the previous section. In particular, the measure focuses upon the concentration of a dataset, rather than its degree of coverage, as with most other diversity measures that have been described in the literature. However, the availability of a linear-time algorithm enables the measure to be applied to many databases, each containing large numbers of molecules, at minimal computational cost, using any type of vectorial representation of a molecule. It might hence be used to provide a rapid way of screening several external datasets to identify some small number that can then be analyzed by more discriminating measures of diversity that take account of factors such as cost, sample availability, and synthetic feasibility, *inter alia*.

ACKNOWLEDGMENT

We thank the following: the Biotechnology and Biological Sciences Research Council and the Engineering and Physical Sciences Research Council for funding; BioByte Corp., Derwent Publications, Val Gillet, and Tripos Inc. for providing the databases used in this study; and Tripos Inc. for software support. The Krebs Institute for Biomolecular Research is a designated Biomolecular Sciences Centre of the Biotechnology and Biological Sciences Research Council.

REFERENCES AND NOTES

(1) Ash, J. E.; Warr, W. A.; Willett, P. *Chemical Information Systems*; Ellis Horwood: Chichester, 1991.

(2) 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.

(3) Shemetulskis, N. E.; Dunbar, J. B.; Dunbar, B. W.; Moreland, D. W.; Humblet, C. Enhancing the Diversity of a Corporate Database Using Chemical Database Clustering and Analysis. *J. Comput.-Aid. Mol. Design* **1995**, *9*, 407−416.

(4) Boyd, S. M.; Beverley, M.; Norskov, L.; Hubbard, R. E. Characterising the Geometric Diversity of Functional Groups in Chemical Databases. *J. Comput.-Aid. Mol. Design* **1995**, *9*, 417−424.

(5) Ashton, M. J.; Jaye, M. C.; Mason, J. S. New Perspectives in Lead Generation II: Evaluating Molecular Diversity. *Drug Discovery Today* **1996**, *1*, 71−78.

(6) Martin, Y. C.; Brown, R. D.; Bures, M. G. In *Combinatorial Chemistry and Molecular Diversity in Drug Design*; Gordon, E. M., Kerwin, J. F., Eds.; (in press).
(7) 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.
(8) Downs, G. M.; Willett, P. Similarity Searching in Databases of Chemical Structures. *Rev. Comput. Chem.* **1995**, *7*, 1−66.
(9) Nilakantan, R.; Bauman, N.; Venkataraghavan, R. A New Method for Rapid Characterisation of Molecular Shape: Applications in Drug Design. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 79−85.
(10) Willett, P. *Similarity and Clustering in Chemical Information Systems*; Research Studies Press: Letchworth, 1987.
(11) Basak, S. C.; Magnuson, V. R.; Regal, R. R. Determining Structural Similarity of Chemicals Using Graph-Theoretic Indices. *Discrete App. Math.* **1988**, *19*, 17−44.
(12) Basak, S. C.; Grunwald, G. D. Tolerance Space and Molecular Similarity. *SAR QSAR Environ. Res.* **1995**, *3*, 265−277.
(13) Downs, G. M.; Willett, P.; Fisanick, W. Similarity Searching and Clustering of Chemical Structure Databases Using Molecular Property Data. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1094−1102.
(14) Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. Chemical Similarity Using Physicochemical Property Descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118−127.
(15) Voorhees, E. M. Implementing Agglomerative Hierarchic Clustering Algorithms for Use in Document Retrieval. *Inf. Proc. Manag.* **1986**, *22*, 465−476.
(16) Holliday, J. D.; Ranade, S. S.; Willett, P. A Fast Algorithm for Selecting Sets of Dissimilar Structures from Large Chemical Databases. *Quant. Struct.-Act. Relat.* **1996**, *14*, 501−506.
(17) Hu, C.-Y.; Xu, L. On Highly Discriminating Molecular Topological Index. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 82−90.

CI960463H