Small Molecule Shape-Fingerprints

James A. Haigh* and Barry T. Pickup

Department of Chemistry, The University of Sheffield, Sheffield S3 7HF, England

J. Andrew Grant

AstraZeneca Pharmaceuticals, Mereside, Macclesfield, Cheshire SK10 4TF, England

Anthony Nicholls

OpenEye Scientific Software Inc., 3600 Cerrillos Road Suite 1107, Santa Fe, New Mexico 87507

Received November 19, 2004

The optimal overlap between two molecular structures is a useful measure of shape similarity. However, it usually requires significant computation. This work describes the design of shape-fingerprints: binary bit strings that encode molecular shape. Standard measures of similarity between two shape-fingerprints are shown to be an excellent surrogate for similarity based on volume overlap but several orders of magnitude faster to compute. Consequently, shape-fingerprints can be used for clustering of large data sets, evaluating the diversity of compound libraries, as descriptors in SAR and as a prescreen for exact shape comparison against large virtual databases. Our results show that a small set of shapes can be used to build these fingerprints and that this set can be applied universally.

INTRODUCTION

Organizing molecules by shape is useful in the lead identification phase of pharmaceutical research. Typically, small subsets of molecules are selected from large collections based on their shape-similarity to compounds of known activity. Experimental screening of such subsets can lead to the direct identification of active compounds.[1] The premise of this approach is that biological receptor sites impose a degree of shape selection, so molecules possessing significant shape similarity can generate a similar pharmacological response. It helps to have the structure of the protein−ligand complex, especially in lead optimization; however, shape matching can be effective with only the structure of an active compound or even with just activity measurements.[2]

Binary fingerprints of molecular properties are used extensively in similarity searching. The bits may represent the presence or absence of a characteristic or whether a numeric property is greater or less than a threshold value. Comparison of bit strings requires only simple Boolean operations allowing thousands, even millions, of comparisons per second. This speed facilitates clustering and neighborhood analysis of very large data sets, for example those produced by High Throughput Screening (HTS). Many molecular descriptors and algorithms have been used to set bit patterns.[3,4] The most commonly used fingerprints contain molecular connectivity information, such as Daylight[5] and Unity fingerprints[6] and MACCS keys.[7,8]

Attempts to describe three-dimensional molecular structure as a fingerprint have previously been based on the presence or absence of pharmacophore patterns and the distances between them.[9−11] Issues with such an approach include the fragility of 'binning' based on intrapharmacophore distances, the arbitrariness of pharmacophore definition and weighting, the lengths of the fingerprints required, and the lack of steric representation.

This paper will describe the performance of shape-fingerprints designed as an excellent approximation to a rigorous measure of the shape overlap. The precise measure of shape-similarity is computed using an efficient Gaussian description of molecular shape[12] to obtain the maximal intersection of the volume of two molecules. The theoretical details of this technique have been described elsewhere,[13] and an implementation of this approach can be found in the program ROCS (Rapid Overlay of Chemical Structures).[1,14]

The ROCS code is optimized to allow approximately 500 Gaussian shape comparisons per second on a single processor. Distributed over many processors it has been used for searching databases of over a billion conformers.[15] However, because there is an initial setup cost, ROCS works best when there is a single query shape. It is not optimized for all-with-all comparisons that are common in clustering, as is necessary in the analysis of HTS activity data or the analysis of molecular diversity. Shape-fingerprints should provide a novel way to include 3D information into such studies, because both comparisons are faster and the typical methods used already expect 2D bit-strings.

The Gaussian shape overlay of two structures A and B allows the evaluation of a volume intersection, V_{AB}. This can be used to define the Shape Distance (SD):

$$\mathrm{SD} = \sqrt{V_{AA} + V_{BB} - 2{*}V_{AB}} \qquad (1)$$

This measure has specific properties[16] including being a true metric, i.e. it obeys the triangle inequality principle. This provides bounds on how different two shapes can be, if the SD is known to a third shape. Although relatively intuitive,

* Corresponding author e-mail: j.a.haigh@sheffield.ac.uk.

10.1021/ci049651v CCC: $30.25
Published on Web 04/19/2005

the formal consequences of this property underpin the strength of the results presented here.

In mathematical terms, a metric invokes a topology on a set of objects. It implies that each shape has a neighborhood and that these are continuously connected. As a result, this 'shape-space' can be represented by reference points whose neighborhoods cover the space. An arbitrary molecule is 'localized' in this shape-space by computing its similarity to these references. This local position represents the absolute shape of a molecule, and these ideas form the basis of the shape-fingerprint. It also implies that the shape of a molecule can be represented as a point in a metric space that has been shown to be non-Euclidean but appears to be of an effectively finite dimensionality.[16] The results presented support this conclusion.

Individual bits of our shape-fingerprint represent 'reference shapes'. Bits are set when the shape similarity measured to the reference shapes is greater than a certain value. The association between a bit and a reference shape is analogous to 'keyed' representations of fingerprints, as opposed to 'hashed' representations where bit positions cannot be associated with specific features.[3] The source of reference shapes and the scheme for transforming continuous values into a binary assignment are described in the *Methods* section. The measures of accuracy and performance of shape-fingerprints relative to numerical optimization are outlined in the *Results* section.

METHODS

Shape-Fingerprints. The Gaussian shape overlay assesses the difference in shape between two molecules by optimizing their Gaussian overlap volume (V_{AB}).[13] This is normalized by using the self-overlap volumes (V_{AA}, V_{BB}) to define the Shape-Tanimoto S_{AB}:

$$S_{AB} = \frac{V_{AB}}{V_{AA} + V_{BB} - V_{AB}} \tag{2}$$

The reference shapes that correspond to bits in the shape-fingerprint are generated by a simple clustering algorithm described in the next section. Bits of the shape-fingerprint are set 'on' if the Shape-Tanimoto to the reference shape exceeds a threshold, termed the 'Bit-On' value. A high Bit-On value results in very few bits being turned on, and consequently a fingerprint can contain too little information to describe the shape accurately. Conversely, a low Bit-On value results in many bits being set, and the saturation can produce a fingerprint with poor discriminatory power. Appropriate settings for Bit-On values are discussed in the *Results* section.

Measures of shape similarity can be obtained from the shape-fingerprints by bit comparison. Such a measure, the shape-fingerprint Tanimoto (SFT_{AB}), analogous to the Shape-Tanimoto S_{AB}, is defined as

$$SFT_{AB} = \frac{N_{AB}}{N_A + N_B - N_{AB}} \tag{3}$$

where N_A and N_B are the number of bits turned on in shape-fingerprints A and B, respectively, and N_{AB} is the number of bits in common to A and B. The parameters and properties of the shape-fingerprint are determined by comparison with Gaussian overlay calculations. Approximating the distance S_{AB} by shape-fingerprints produces a measure (SFT_{AB}) that most accurately describes short distances (similar molecules) but discriminates poorly between dissimilar molecules. For most clustering or searching applications this is not a problem, because precise information is only required about similar molecules, and information pertaining to distances beyond a certain threshold is discarded. It is not appropriate, therefore, to assess the accuracy of SFT_{AB} using linear regression on the entire range of values of S_{AB}.

Instead, a method was used to quantify the performance of SFT_{AB} that emphasizes the correctness of describing very similar molecules. In this approach, the similarity between a query and each molecule in a database is computed using shape-fingerprints. This generates an ordered list of molecules according to SFT_{AB}. To assess the accuracy of the fingerprint, the Gaussian overlays are computed for the query molecule against the same database, to determine the most similar N molecules. These are referred to as the 'ideal retrieval set'. The value of N here is typically 1000 as this is the upper end of the number that would be visually examined, and the lower end of the number of compounds used in an accurate secondary-screening assay. Receiver Operator Characteristic (ROC) curves are used to quantify the performance of shape-fingerprints in correctly identifying the ideal retrieval set.

ROC Curves. ROC curves, widely used in signal detection and medical statistics,[17,18] depict the relative tradeoff between success and failure rates. ROC curves fit the requirements for a measure that emphasizes correct prediction of a small subset. For instance, their origins lie in radar detection where they were used to correctly spot real events in noisy data. Similarly in medicine the majority of subjects are healthy, but those with a condition need to be discerned. In shape analysis, we are typically looking for at most 10%, possibly as low as 0.1% of the total available structures.

ROC curves are produced by classifying data as positive or negative according to a threshold decision. Each position in the ordered list of shapes defines a threshold by its SFT_{AB} to the query. Molecules above the threshold found in the ideal retrieval set are defined to be the true positives (TP); those not are the false positives (FP); molecules below the threshold and in the ideal retrieval set are the false negatives (FN); remaining molecules are true negatives (TN).

The ROC curve is a plot of the true positive rate (TPR) on the y-axis against the false positive rate (FPR) on the x-axis, calculated at intervals over the ordered list. The TPR and FPR are normalized to vary between zero and one and are defined as follows:

$$TPR = \frac{TP}{TP + FN} \tag{4}$$

$$FPR = \frac{FP}{TN + FP} \tag{5}$$

A good shape-fingerprint search ranks the true positives higher than true negatives, having a TPR greater than the FPR. This ROC curve is generated by assessing the TPR and FPR at different fractions of the list, ordered by fingerprint Tanimoto, or, equivalently at different threshold

Tanimoto values. A perfect search produces a ROC curve passing through the point (0,1), whereas a shape-fingerprint with no discriminatory power to rank true positives results in a line with unit slope, i.e. as would a randomized list. Note that it is possible for a ROC curve to lie below the unit slope, in which case an unexpected negative correlation has been identified.

A different method of appraising the facility of a method to extract 'needles-from-a-haystack' is to calculate 'enrichments'. These are found by comparing the fraction of actives in the top P percent of the list against the whole list. The issue with enrichments is that they vary with P. If the top structure is an active, choosing P as $100/N$ (where N is the number of molecules) gives a perfect enrichment but not a very useful one—for instance, the next active might be at the bottom of the list. The advantage of ROC curves is that there is a measure that is independent of threshold or percentage. The area under the curve (AUC) represents all thresholds and is a measure of the information content of the characteristic, or score, used to estimate activity. A perfect ROC curve has an AUC of 1.0; a random result has a unit slope ROC curve, and an AUC of 0.5. Experience shows that an AUC of 0.9 or greater is an indication of a useful measure. At around 0.95 a typical estimation of the enrichment is about a factor of 10.

Reference Shape Generation Algorithm. The algorithm used to generate diverse reference shapes also captures the features of the shape-space of those molecules. It is based on the standard hierarchical clustering 'dmax' approach[19] and depends on a single parameter, the Design-Tanimoto, used as a threshold value for deciding if a shape belongs to a given cluster. Working from a database of structures $S = \{s_1, s_2 .... s_N\}$, the algorithm generates a set of reference shapes $R = \{r_1, r_2 .... r_M\}$. The first reference shape r_1 can be chosen at random. For each member of S the Shape-Tanimoto with respect to r_1, ST(s_n, r_1), is compared to the Design-Tanimoto. If ST(s_n, r_1) is above the Design-Tanimoto, the shape s_n is close to r_1, and it is assigned to this reference shape and plays no further part. Otherwise, ST(s_n, r_1) is stored in the closest reference list (CRL). After all structures in S have been tested, the shape that is least similar to r_1, i.e. with the smallest value in the CRL, is chosen as r_2. Shape-Tanimotos are computed between r_2 and the remaining members of S. Members of S are assigned to r_2 if the Shape-Tanimoto exceeds the Design-Tanimoto. Furthermore, the CRL is updated so that for each unassigned member of S, it stores the value of the Shape-Tanimoto to the most similar shape in R. For each iteration i, members of S are assigned to r_i, the CRL updated, and a new reference shape r_{i+1} selected as before. The algorithm terminates when all members of S have been assigned to a reference. The result is a set of reference shapes distributed across shape-space such that no pair of reference shapes has a similarity greater than the Design-Tanimoto. Increasing the Design-Tanimoto threshold produces more reference shapes and a more detailed description of shape-space; decreasing it produces a coarser, simpler description.

RESULTS

Reference Shape Generation. The databases used for all calculations are from the Cambridge Structural Database (CSD)[20,21] and the MDL Drug Data Report (MDDR).[22] A subset of organic molecules with a molecular weight range of 120−500 was chosen from the CSD. The resulting set contains 46 182 experimental structures, with an average molecular weight of 278. A second subset of 50 000 molecules was chosen from the MDDR that satisfied the following criteria: molecular weight range 100−750, ClogP[23] range {−6.0, 6.0}, and fewer than 11 rotatable bonds. The structures were transformed into consistent tautomeric forms, and 3D structures were generated using Corina.[24] This subset has an average molecular weight of 375.

Table 1. Number of Reference Shapes Generated with Increasing Design-Tanimoto for All Molecules[a]

Design-Tanimoto	CSD reference shapes	MDDR reference shapes
0.50	23 (15)	40 (12)
0.55	50 (38)	120 (26)
0.60	136 (88)	339 (77)
0.65	393 (252)	987 (237)
0.70	1150 (820)	2511 (837)
0.75	3119 (2473)	5509 (2684)
0.80	7173 (6353)	10567 (6660)
0.85	13752 (12730)	17694 (12906)

[a] Molecules with 12−32 heavy atoms are given in parentheses.

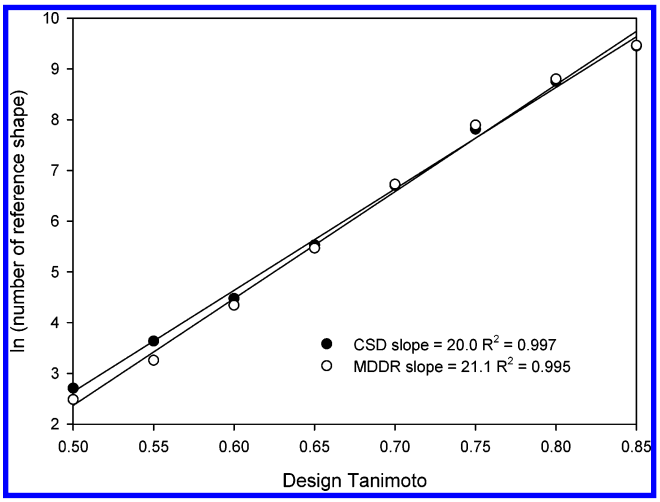

Figure 1. Variation of the number of reference shapes with Design-Tanimoto, for different sets of molecules with 12−32 heavy atoms.

Table 1 shows the exponential increase in the number of reference shapes (hence length of shape-fingerprint) for different Design-Tanimoto values. At any Design-Tanimoto the MDDR molecule subset has the higher number of reference shapes. Since the two sets had essentially the same number of total shapes, this suggested a more diverse range of shapes in the MDDR molecule subset. To test whether this arose from the greater molecular weight range in the MDDR set, subsets of each data set were produced with a heavy atom count of between 12 and 32. The CSD was thus reduced to 40 524 molecules, and the MDDR to 40 659. The number of reference shapes for these sets are given in parentheses in Table 1 and graphed in Figure 1. While it is true that the Corina algorithm used to generate the MDDR structures uses torsional potentials and ring templates based on parameters obtained from the experimental CSD structures, the two progressions are still remarkably similar, suggesting that the collections occupy very similar shape spaces.

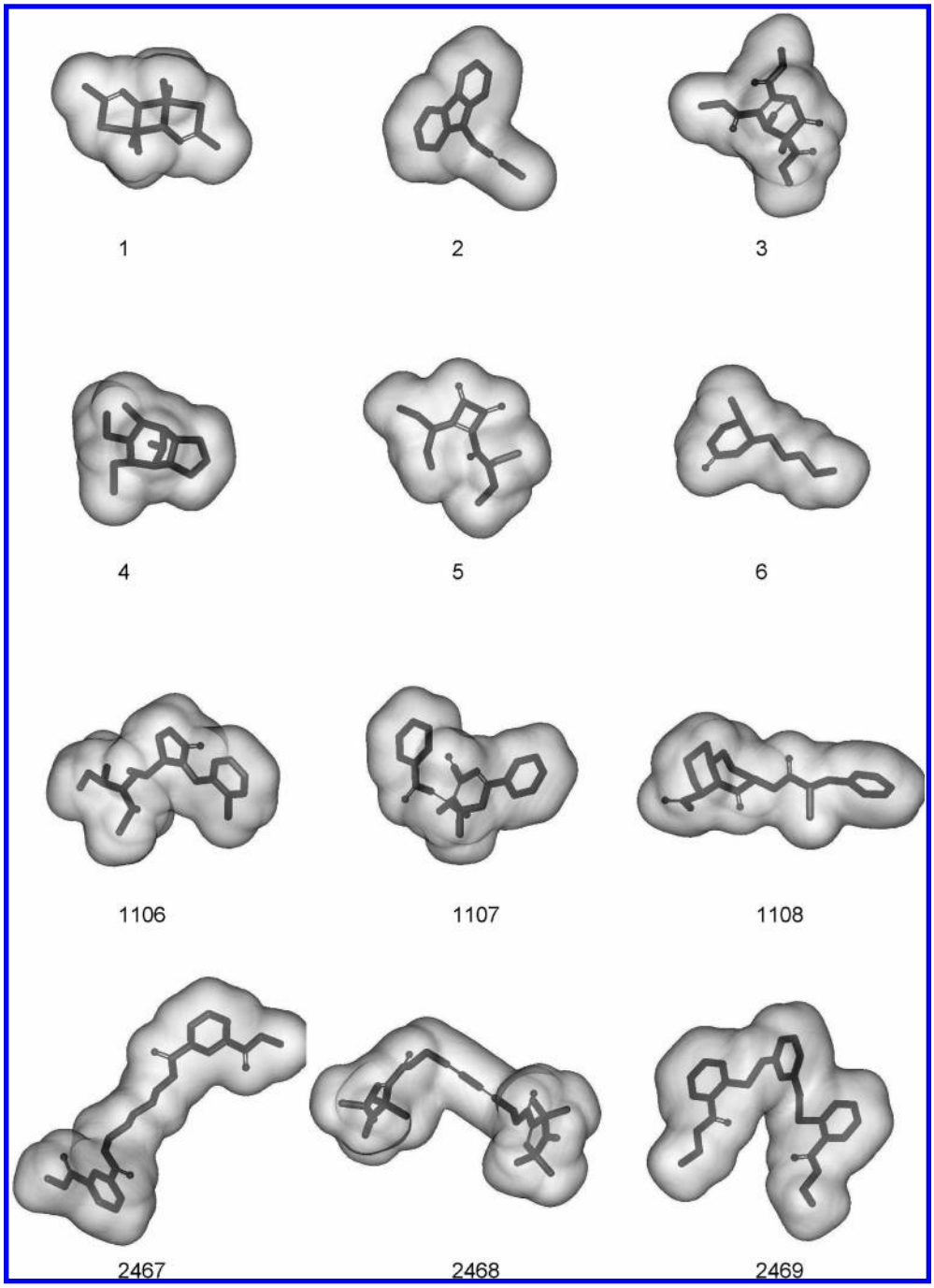

Figure 2. Examples of CSD reference shapes obtained using a Design-Tanimoto of 0.75. They are ordered in terms of decreasing likelihood of an arbitrary structure being within the neighborhood of each shape. (See text for more details.)

Some examples of reference shapes are shown in Figure 2. These shapes are obtained from the CSD heavy atom limited set after clustering at a Design-Tanimoto of 0.75. The shapes have been ordered by the likelihood that the bits they represent are set 'on' in a fingerprint. The likelihood is estimated from analysis of a large (5.2 million) set of shape-fingerprints. The index number associated with each shape in the figure indicates its popularity, where the shape labeled 1 is the most common.

The procedure was also applied to multiconformer data sets. Two databases were generated comprising 8500 and 55 000 randomly chosen commercially available molecules (CNC) with a heavy atom count of between 12 and 32. Corina[24] structures were obtained and expanded into conformers using Omega.[25] An Omega parameter controlling the range of conformational energies was adjusted to produce different multiconformer databases for both of the sets with between 50K and 500K conformers. As shown in Figure 3, the same straight-line behavior is seen between Design-Tanimoto values of 0.5 and 0.75. The slopes of these lines correlate with the total number of structures in each data set and are almost independent of the origin of the structures, i.e. whether single-conformer or multiconformer, or the degree of multiconformational expansion.

Fingerprint Performance. Figure 5 shows ROC curves for fingerprint searches of the CSD database using a query molecule from the CSD, designated by the code PEWHII, as a function of Bit-On values. The shape of this molecule, and those others referenced explicitly by CSD code in the Results section, is illustrated in Figure 4. The reference

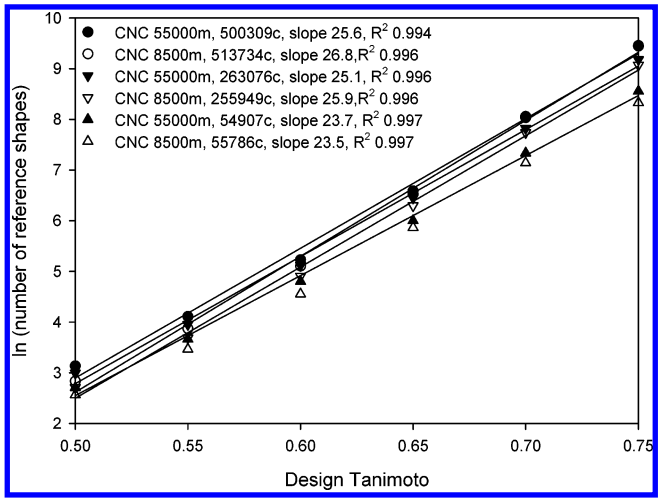

Figure 3. Variation of the number of reference shapes with Design-Tanimoto, for different multiconformer databases, each with 12−32 heavy atoms. (m and c abbreviate molecules and conformers, respectively.)

shapes were generated with a Design-Tanimoto value of 0.75, producing a shape-fingerprint of 3119 bits. Gaussian shape-overlay calculations are carried out to determine the 1000 most similar molecules to the query molecule, the set being referred to as the ideal retrieval set. A Bit-On value equal to the Design-Tanimoto (in this case 0.75) produces a ROC curve along the diagonal ($y = x$). This indicates that such fingerprints perform randomly in terms of correctly discriminating between the true positives (shape matches from the ideal retrieval set) and true negatives. Setting the Bit-On value lower than the Design-Tanimoto generates a shape-fingerprint that gives ROC curves with a high TPR relative to the FPR. The optimal Bit-On value is 0.65, and approximately 90% of the correct shape-neighbors of PEWHII are identified with few misclassifications (FPR $\approx$ 10%).

Figure 6 shows ROC curves for a query molecule also from the CSD, with the code HIFYOK. Setting the Bit-On value to equal the Design-Tanimoto again produces a fingerprint with no ability to discern positive from negative instances. However, this graph shows a different behavior than that for PEWHII. In Figure 6 the optimal Bit-On value is also 0.65 but does not appear quite as efficient at identifying true positives, and the curves for other Bit-On values are worse than in Figure 5.

It can be seen from Figures 5 and 6 that the ROC curves show a clear variation in search performance depending on the choice of Bit-On value and the query molecule. Differences between the curves can be quantified by computing the area under the curve (AUC), which varies between 0.50 (random) and 1.00 (perfect).

Table 2. Analysis of AUC Distributions Using a Fingerprint of 3119 CSD Reference Shapes, Obtained Using a Design-Tanimoto of 0.75[a]

Bit-On value	0.55	0.60	0.65	0.70	0.75	0.80
mean AUC	0.925	0.956	0.973	0.948	0.753	0.546
standard deviation of AUC	0.046	0.026	0.025	0.090	0.177	0.110

[a] The distributions of Bit-On 0.65 and 0.75 are shown in Figure 5.

The variation in search performance is reflected in the range of AUC values computed for Bit-On values below 0.75. For the query PEWHII this range is 0.92−0.98, whereas for the poorer ROC curves associated with HIFYOK the range of AUC is 0.86−0.95. To obtain better statistics of the variation in AUC, 1000 molecules were chosen at random from the CSD and MDDR databases to define queries used in the ROC analysis. Random selection ensures that there is neither bias toward molecules containing many shape neighbors nor intentional correlation between the shapes of the queries. These two query sets are used in the following sections that investigate various properties and design features of the shape-fingerprint.

Analysis of Bit-On Value. This section investigates in more detail the result of variation of the Bit-On value using the two large sets of query molecules and describes an approach to obtain optimal Bit-On values for a range of fingerprint lengths.

Figure 7 shows the distributions of AUC for Bit-On values of 0.65 and 0.75 computed for each of the 1K CSD query molecules using the fingerprint obtained from a Design-Tanimoto of 0.75. The Bit-On value of 0.65 results in a sharp distribution with a high mean area (0.97) and suggests a good search method. The Bit-On of 0.75 shows variable performance, with a mean area halfway between random and perfect discrimination. Table 2 shows the mean (μ) and standard deviations (σ) in these distributions of AUC over a range of Bit-On values. The best values of μ and σ are obtained for a value of 0.65, while the poorest performance is when the Bit-On value equals or exceeds the Design-Tanimoto.

Tables 3 and 4 show results for MDDR fingerprints of different lengths corresponding to different Design-Tanimoto values. Both show the same trends as for the CSD, with a peak performance of the Bit-On value at around 0.60−0.65, and a drop-off when the value equals or exceeds the Design-Tanimoto.

In Figure 7, at the optimal Bit-On of 0.65, there are approximately 20 outliers in the distribution with areas less than 0.90. These were suspected to correspond to query molecules with few close shape-neighbors in the database. Consequently, when searching for 1000 neighbors, a sig-

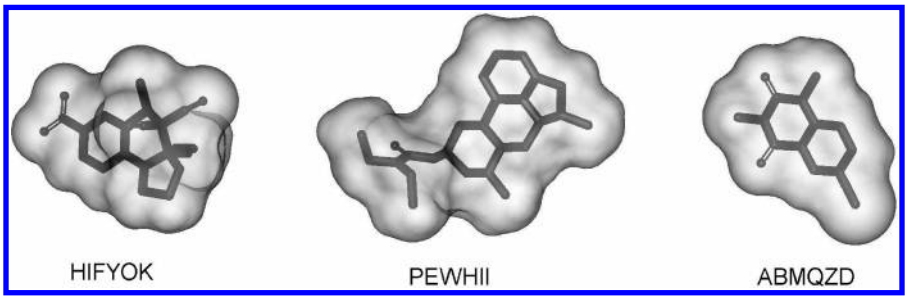

Figure 4. Illustration of the shape of those molecules referred to by CSD code in the Results section.

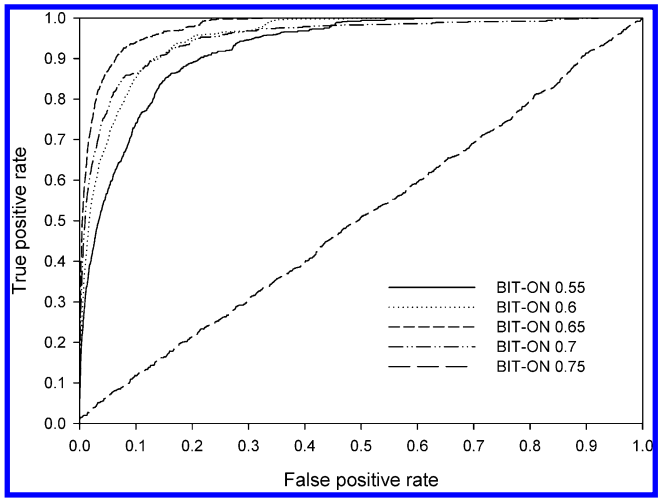

Figure 5. ROC curves for molecule PEWHII, searching for 1000 nearest shape neighbors in the CSD, using fingerprints generated from 3119 CSD reference shapes.

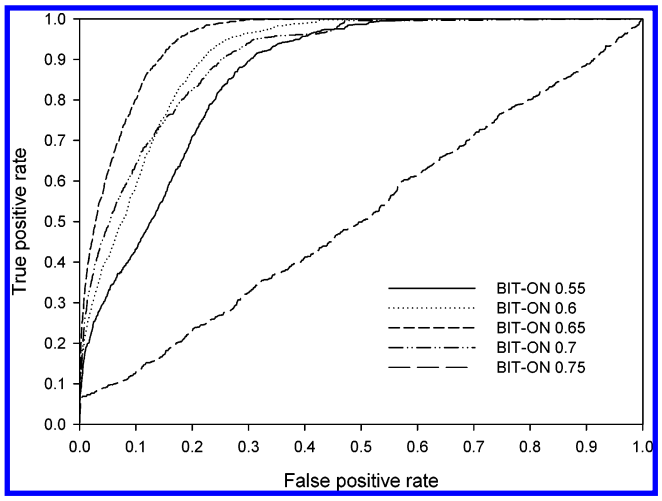

Figure 6. ROC curves for molecule HIFYOK, searching for 1000 nearest shape neighbors in the CSD, using fingerprints generated from 3119 CSD reference shapes.

nificant number would have small values of S_{AB}. An S_{AB} below 0.7 corresponds to a low level of visible molecular similarity that the shape-fingerprint is not designed to quantify. A simple measure of the number of nearby shape-neighbors a query molecule has in a database is the Shape-Tanimoto of the last neighbor in the ideal retrieval set (in these graphs the Shape-Tanimoto of the 1000th neighbor). Typically, molecules with low values have fewer close neighbors. Figure 8 shows the variation in the AUC with the value of the Shape-Tanimoto of the 1000th neighbor for the data set in Figure 1. Clearly query molecules with low AUC values have the most distant shape neighbors. In the studied set none of the queries for which the AUC is less than 0.90 had 1000 shape-neighbors, out of more than 40 000 molecules, which would be considered visibly similar.

Figure 9 compares the results obtained at Bit-On values of 0.60 and 0.70 with the optimal value of 0.65. Although the Bit-On value of 0.60 improves the results on the low-scale of the *x*-axis, there is considerable degradation in the AUC on the high end. Conversely, 0.70 gives poor AUC on the low-scale and barely discernible improvement at the high-end. This graphical analysis suggests an optimal Bit-On value consistent with the data in Table 2.

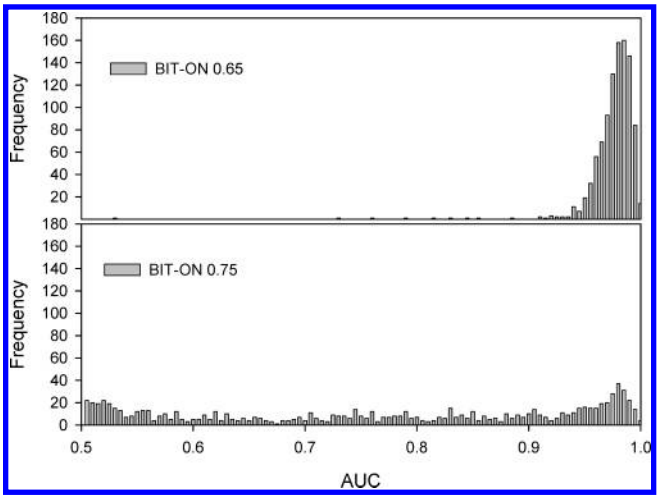

Figure 7. Distribution of AUC for Bit-On values 0.65 and 0.75 computed for 1000 query shapes from the CSD. Each AUC represents the performance of searching for 1000 shape-neighbors of each query in the CSD. The fingerprint contains 3119 CSD reference shapes, obtained using a Design-Tanimoto of 0.75.

Table 3. Analysis of AUC Distribution Using a Fingerprint of 5509 MDDR Reference Shapes Obtained Using a Design-Tanimoto of 0.75

Bit-On value	0.55	0.60	0.65	0.70	0.75	0.80
mean AUC	0.959	0.975	0.972	0.892	0.648	0.505
standard deviation of AUC	0.019	0.015	0.048	0.145	0.147	0.048

Table 4. Analysis of AUC Distribution Using a Fingerprint of 2511 MDDR Reference Shapes Obtained Using a Design-Tanimoto of 0.70

Bit-On value	0.55	0.60	0.65	0.70	0.75	0.80
mean AUC	0.964	0.975	0.941	0.741	0.741	0.496
standard deviation of AUC	0.017	0.018	0.084	0.161	0.088	0.044

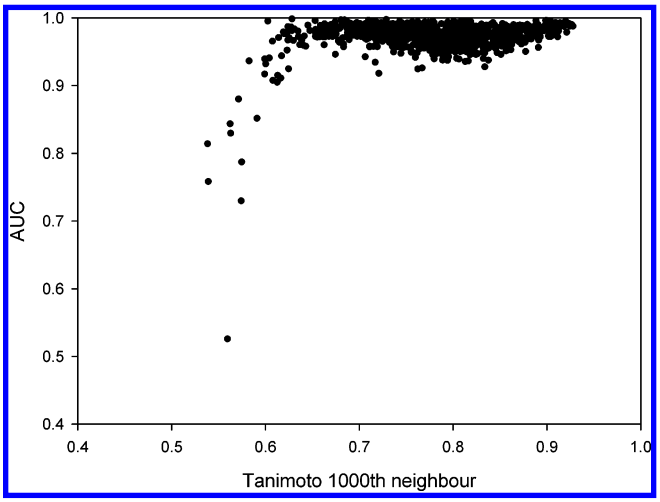

Figure 8. Variation of the AUC with the Shape-Tanimoto of the last neighbor (1000th) in the ideal retrieval set, for the 1K CSD set, using a Bit-On of 0.65, and a fingerprint constructed from 3119 CSD reference shapes.

This analysis has been performed on the shape-fingerprints of varying lengths, for both the CSD and the MDDR, each showing similar behavior. There is always a single Bit-On value that produces the best results, and this value lies below the corresponding Design-Tanimoto value. This optimal Bit-On threshold approximately determines the range of applicability of the SFT_{AB}. Values of S_{AB} greater that the Bit-

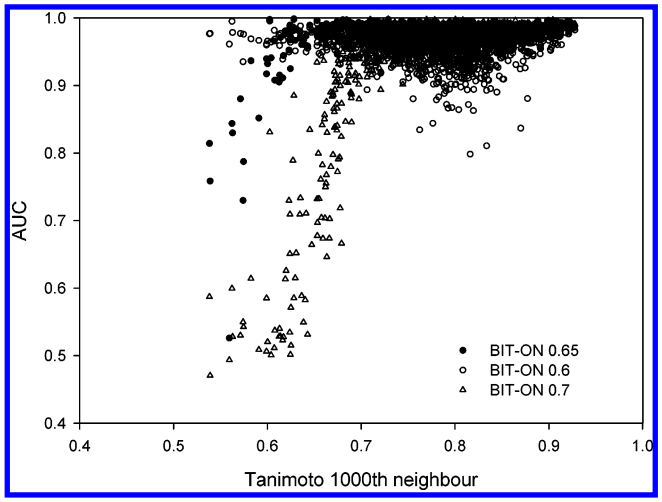

Figure 9. Variation of the AUC with the Shape-Tanimoto of the last neighbor (1000th) in the ideal retrieval set, for the 1K CSD set, using different Bit-On values and a fingerprint constructed from 3119 CSD reference shapes.

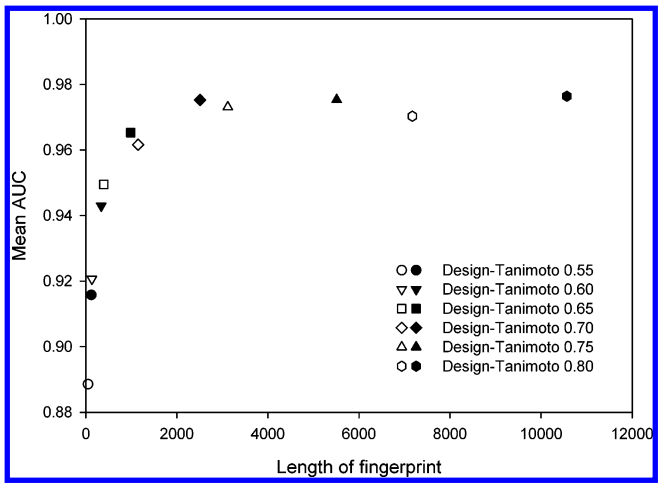

Figure 10. Variation in the mean AUC as a function of shape-fingerprint length. The calculations are for CSD (open symbols) and MDDR (closed symbols) data sets (see text). The optimal Bit-On value is used for each fingerprint.

On value correlate well with SFT_{AB}, whereas values of S_{AB} smaller than the Bit-On value correlate poorly.

Fingerprint Length. The impact of adding more bits to the shape-fingerprint is shown in Figure 10, a plot of shape-fingerprint length against the variation in the mean AUC, at the optimal Bit-On value, for the 1K CSD and MDDR query sets. These fingerprints are constructed from reference shapes obtained from clustering at Design-Tanimotos in the range 0.55−0.8.

It might be expected that adding extra bits (i.e. reference shapes) would improve their information content and hence the ability to find shape-similar molecules. Indeed, in Figure 10, the performance, as measured by the mean AUC, does increase from the smallest lengths to an optimal value of between 3000 and 4000. Beyond this point additional reference shapes do not seem to add information. The relatively small length of the optimal fingerprint is comparable to other binary representations of 2D descriptor information. Even at fingerprint lengths of 500−1000, significant shape information is preserved. The convergence to an optimal length from different data sets suggests that the shapes of molecules belong to an effectively finite, universal space. This hypothesis is examined in the next section.

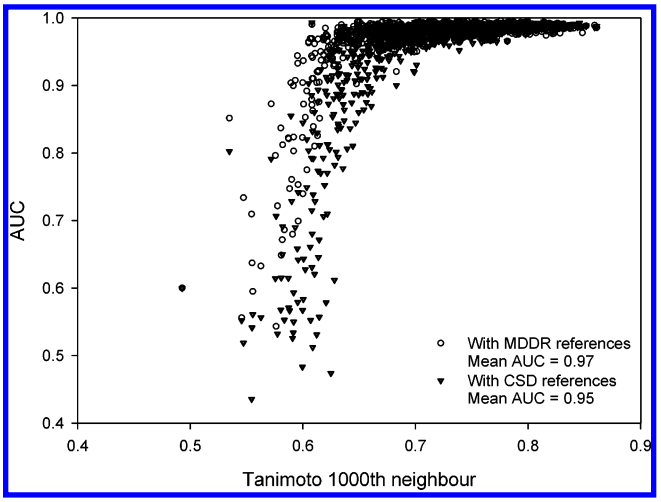

Figure 11. Variation of the AUC with the Shape-Tanimoto of the last neighbor (1000th) in the ideal retrieval set, for the 1K MDDR set, using a Bit-On value of 0.65. Comparison is made between fingerprints constructed from 5509 MDDR and 3119 CSD reference shapes, to search the 50K MDDR data set.

Transferability. The advantages of transferable shape-fingerprints, i.e. reference shapes from one set of molecules used for comparison within another, would be considerable. It would be unnecessary to derive a separate set of reference shapes for each data set analyzed. The shape diversity of a given set of compounds could be evaluated with the knowledge that there are representatives of all likely shapes encoded in each fingerprint.

To test the hypothesis of universality, 1000 shapes searches, using the previously randomly selected MDDR queries, were performed but with fingerprints generated with reference structures from the CSD. Figure 11 compares these results (triangles) with similarity searches with fingerprints constructed from MDDR reference structures (circles). A Bit-On value of 0.65 was used. These plots of the AUC versus the Shape-Tanimoto of the 1000th neighbor have the same characteristics as described for the analysis of the CSD database with CSD fingerprints. There is little difference in the value of the AUC computed using either set of reference shapes to obtain shape-fingerprints. The largest discrepancies occur for query molecules with the fewest shape neighbors. Considering only points in which the 1000th Tanimoto value is 0.7 or higher, there is an insignificant drop in the mean AUC (∼0.002). Figure 12 shows the results for the same query set and reference shapes but with 0.6 as the Bit-On value. The graph shows the same features as the Bit-On 0.65 data (Figure 11). The main difference in behavior of the two Bit-On values is that at 0.6 the average performance, as reflected by the mean AUC, is slightly improved over 0.65. This is due to the significant improvement of the outliers, molecules that have low values of the Shape-Tanimoto of 1000th neighbor. However, this overall improvement obscures the fact that the performance on molecules that have high values of the Shape-Tanimoto of 1000th neighbor is not as good as at 0.65.

Figure 13 shows the performance of using the approach for a database of half a million commercially available compounds containing 12−32 heavy atoms, using the 0.75 Design-Tanimoto CSD reference shapes. The size of this

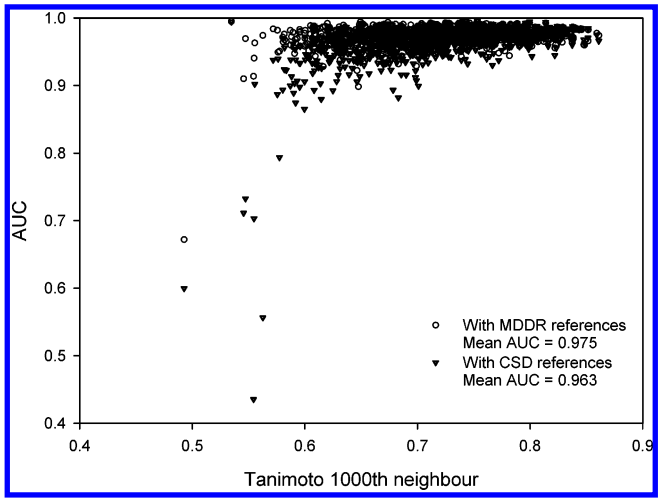

Figure 12. Variation of the AUC with the Shape-Tanimoto of the last neighbor (1000th) in the ideal retrieval set, for the 1K MDDR set, using a Bit-On value of 0.60. Comparison is made as in Figure 9.

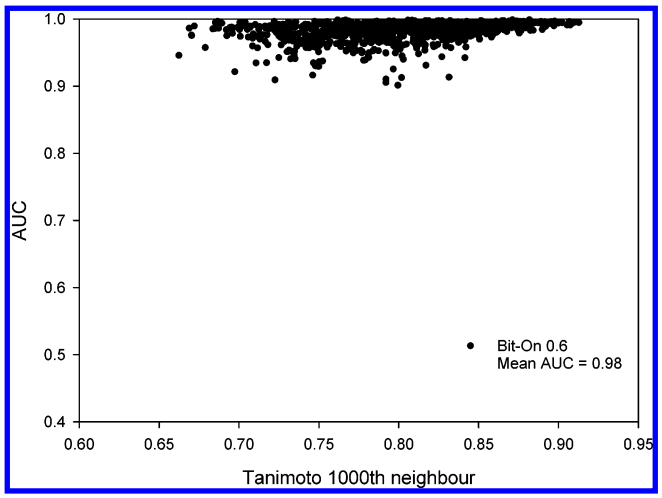

Figure 13. Variation of the AUC with the Shape-Tanimoto of the last neighbor (1000th) in the ideal retrieval set, for 500 000 commercially available molecules containing 12−32 heavy atoms. The fingerprints are constructed from 2473 CSD reference shapes also containing 12−32 heavy atoms.

database is typical of those used in drug-discovery screening assays. The queries are a 1000 molecule random subset chosen from this database. These results also demonstrate the transferability of the independently derived set of CSD reference shapes. These calculations had no significant bad outliers using a Bit-On value of 0.60, whereas for a Bit-On value of 0.65 there are a small number of searches that produce an AUC less than 0.90.

Properties of Shape-Fingerprints. The 'bit density' of a fingerprint is defined as the ratio of the number of bits turned on to the total number of bits in the fingerprint. It is a characteristic property of fingerprints related to their power of discrimination. To obtain statistical information about this property, shape-fingerprints built using the CSD reference shapes (limited to 12−32 heavy atoms) were computed for 5.2 million molecules taken from a set of commercially available compounds. The average bit density is 0.15 (at Bit-On 0.6) and 0.055 (Bit-On 0.65), for a fingerprint of bit length 2473. There is a considerable range in the density, but the upper limit, which is rarely found, is 0.3, and thus shape-fingerprints are far from saturated. Our results confirm

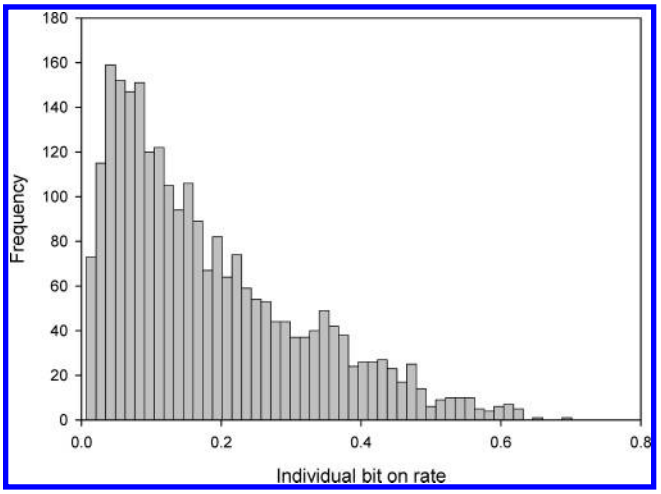

Figure 14. Distribution of counts for bits used in 5.2 million shape-fingerprints, constructed using 2473 CSD reference shapes and a Bit-On value of 0.6. To simplify the presentation of the *x*-axis the counts are normalized by the total number of fingerprints.

that there is sufficient information in the fingerprint for accurate searches. A feature of 2D chemical fingerprints, based on molecular graphs, is that as molecules increase in size and complexity they tend to set more bits, giving rise to saturation effects.[3] By contrast, the shape-fingerprint exhibits a maximum in the bit-density for medium size molecules. This is because such molecules can potentially turn on bits for both slightly larger and slightly smaller molecules and hence have the most 'similarly sized' neighbors.

Each bit in a fingerprint has a given probability of being turned on for a set of structures. Using the 5.2 million fingerprints described above, these probabilities were calculated and binned. The frequency of such probabilities is illustrated in Figure 14. This shows nearly all bits are more often off than on, and none are always on. Most reference shapes bits are turned on with probabilities of less than a few percent. This again is in contrast to 2D fingerprints where some keys have common occurrence, for instance aromatic rings. This analysis was used to generate the orderings of the reference shapes shown in Figure 2. The figure highlights the most common reference shapes as well as those that occur with lower probability.

A 4000-bit fingerprint requires 500 bytes of storage. This is large when compared to typical storage requirements, per conformer, of typical molecular file formats. However, as Figure 14 illustrates, most bits are turned on infrequently, as such fingerprints are dominated by zeros. Using standard computational approaches for compressing binary data, the average storage requirement per shape-fingerprint can be reduced to around 80 bytes. At this level of compression it could easily be stored along with a molecular structure, for instance as tagged data in an MDL SD file.[26]

Comparison to Shape Multipoles. An alternative approach to evaluating shape similarity is to define an intrinsic property describing molecular shape that can be used to make comparisons. A useful method for the characterization of molecular shape is in terms of a set of moment averages or 'shape multipoles'.[27,28] These multipoles can be obtained analytically with trivial cost, and a dimensionless similarity function can be computed based on the difference in the components of the multipoles.[27]

Table 5. Mean (μ) and Standard Deviation (σ) of AUC Distributions Using the Shape Multipole Method[a]

	CSD AUC		MDDR AUC	
	μ	σ	μ	σ
V	0.774	0.090	0.724	0.110
Q	0.885	0.057	0.861	0.063
Q+*O*	0.893	0.054	0.875	0.059

[a] The similarity function uses the difference in volume (*V*), quadrupole (*Q*), and quadrupole combined with octupole (*Q*+*O*).

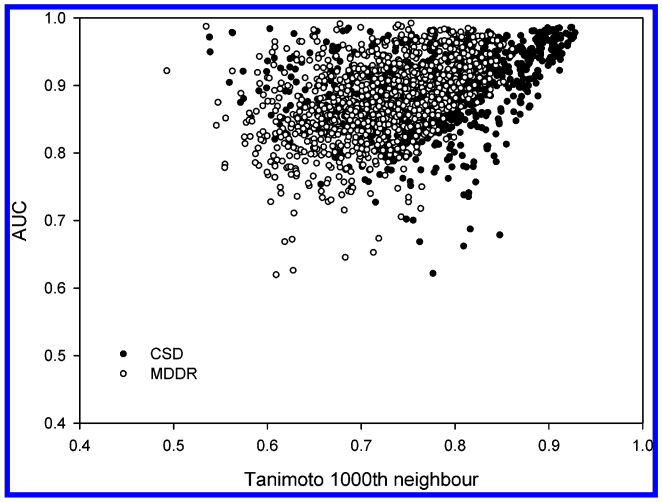

Figure 15. Variation of the AUC with Shape-Tanimoto of the last neighbor (1000th) in the ideal retrieval set, for the 1K CSD and MDDR sets, using multipole similarity (quadrupole combined with octupole).

This approach was used to search the CSD and MDDR databases using the corresponding 1K query sets. The ROC analysis was used to compare these search results with the ideal retrieval sets obtained from using the Gaussian overlay. Table 5 shows the mean and standard deviation in the AUC for the multipole method, in which the similarity function comprised volume, quadrupole, and finally quadrupole combined with octupole. As expected the performance of the search method is very poor if only molecular volume is used to define the similarity. There is improvement if the quadrupole term is considered and a slight further improvement if this term is combined with the octupole. The search performance as characterized by the AUC is clearly not as good as using the shape-fingerprint method with optimal Bit-On values (see Tables 2−4, columns 0.6 and 0.65).

Figure 15 plots the variation in the AUC with the value of the Shape-Tanimoto of the 1000th neighbor of the query molecule. There is no clear correlation between the performance of the search and the shape of the query molecule, differing greatly from the patterns seen with shape-fingerprint method, for instance Figures 9 and 11. Shape-multipoles are less discriminating in distinguishing features of molecular shape. This is illustrated in Figure 16, which plots for the single molecule (ABMQZD, as shown in Figure 4) the multipole similarity versus the Shape-Tanimoto computed with respect to the complete CSD database. At low multipole similarity (<2.0) there are many shape overlays of poor shape matches, in addition to the best shape fits.

DISCUSSION AND FUTURE DIRECTIONS

The results presented here on the properties and use of shape-fingerprints can be understood and, indeed, expected

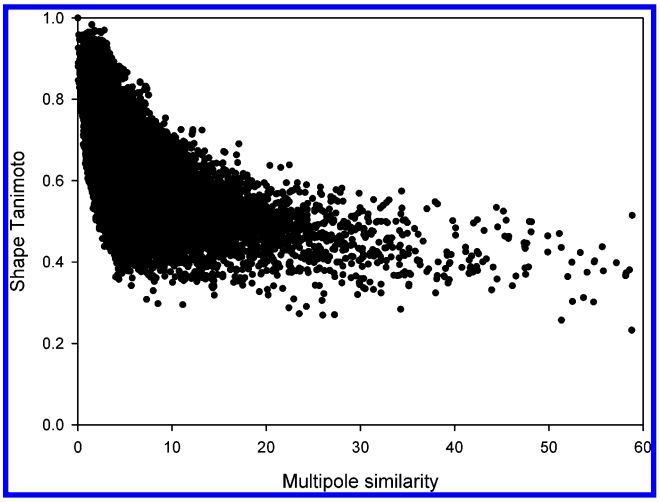

Figure 16. Plot of multipole similarity versus Shape-Tanimoto. The similarities are computed for the molecule ABMQZD with respect to the entire CSD data set.

in the light of the metric properties of molecular shape. A strict metric induces a topology, which implies a space with the property of neighborhood, i.e. one that can be described by representative points. One criticism might be that, given such a powerful property, it would make more sense to exactly deconstruct such a space by traditional techniques and use those results to describe molecular shape. A paper in preparation outlines such an approach. However, there are several reasons for recommending shape-fingerprints as described. For one, it is in the current language of chemical informatics, for which there is considerable computational machinery. It is also considerably less abstract than a more mathematical description. The reference shapes that make up a shape-fingerprint, although not canonical, are familiar enough to medicinal chemistry to provide interpretability. Finally, the mathematics of shape space is far from elementary. Shape does not embed into a Euclidean or finite space.[16] As such, for many applications, fingerprints may turn out to be the more efficient and useful path to the inclusion of a rigorous definition of molecular shape into chemistry.

In this work, constructing a set of representative shapes is equivalent to covering the metric space with a set of hyperspheres of radii equal to one minus the Design-Tanimoto, also known as the Soergel measure.[4] These spheres intersect but their centers are interior only to themselves. Each structure then falls within the radius, defined by one minus the Bit-On value, of one or more of these spheres. This set of local spheres then provides a description of the shape. It is the nature of high-dimensional spaces that local volumes are a steep function of the distance. As such, many neighbors are required to accurately define a position. If a shape has too few neighbors, its position is poorly defined. The opposite problem, too many neighbors, can occur if the Design-Tanimoto is set too low, i.e. the neighborhoods are too large. At the limit, all shapes are in the neighborhood of a few reference shapes and there is no discrimination. One consequence of this observation is that it might be expected that choosing representative structures at random would result in poorly populated neighborhoods being less well represented and conversely the popular neighborhoods to be over-represented, with both effects contributing to a poor performance of fingerprints so generated. Indeed, we have seen this for shape-fingerprints

constructed for CSD and MDDR molecules using randomly chosen reference shapes.[29] They were found to perform systematically worse than those that utilize reference shapes selected to reflect the underlying structure of shape-space. The resulting fingerprints have a higher mean bit density but also a significant fraction have no bits turned on. It is, however, possible that other algorithms might also extract useful sets of references structures and such work is continuing.

Perhaps the most striking aspect of the reference structure design process is the linearity seen in the density plots of Figures 1 and 3. It can be shown that the forms of such curves arise naturally from stochastic selection processes without replacement, i.e. where each successful selection must be different from all prior selections and where the probability of similarity to any such prior selection is a strong function of a 'radius', in this case greater than the fifth power. Not so obvious, but also important, are the two ends of the curve, both of which provide upper limits on the dimension of shape space. Very high dimensional spaces suffer from the property that everything is far away from everything else. This is clearly not true for shape. There are relative few structures at a Design-Tanimoto of 0.5. In addition, the number of reference shapes does not saturate until the Design-Tanimoto is close to unity. A high dimensional space would saturate much earlier. This, and the eigenvalue analysis of classical multidimensional scaling,[16] suggests there are relatively few (between 5 and 10) important dimensions, although more may be required for highly accurate mapping.

It is interesting that although we see some evidence of saturation from the density curves, in general the more structures reference shapes are drawn from, the more are found. However, the radius implied by the optimal Bit-On value is significantly larger than that given by the Design-Tanimoto and so it is not necessary to find all possible spheres, i.e. complete coverage of shape-space. In fact, our results show that the Design-Tanimoto is a bad choice for the Bit-On value: structures are poorly defined by just one shape unless, presumably, the resolution is impractically high. In early work the concept of a single representative structure was explored but found inferior to that presented. Work continues on the asymptotic character of these curves. In particular, we have noticed that multiconformer structures generate more reference structures for a given total of shapes, indicating, as might be expected, that they occupy a slightly expanded shape space compared to the lowest energy shapes.

The comparison of shape-fingerprints to shape multipoles provides an interesting contrast in efficacy. If two structures have the same shape, then they have the same shape multipoles, but if they have the same multipoles, then they do not necessarily have the same shape (as exemplified in Figure 16). Shape multipoles condense shape information, i.e. involve the loss of information. Most, if not all, methods that claim to capture molecular shape have this property; they 'map' many shapes to one descriptor (set). Fingerprint methods in the 2D domain have a similar property; there may be many molecules with the same fingerprint. This is exacerbated by the 'folding' of fingerprints. However, shape-fingerprints are an attempt to precisely map a molecule's position in shape space. As such, similarly shaped molecules should have the same fingerprint, i.e. the correspondence is one-to-one not many-to-one. What prevents a precise mapping is that neighborhoods change rapidly in high-dimensional spaces, i.e. the representative shapes may vary sharply as shape changes. This presumably explains why the optimal Bit-On threshold is significantly lower than the Design-Tanimoto: representative structures that are close to the target shape will vary too rapidly as a function of shape. A broader distribution, or 'cloud', of representatives remains more stable from shape to similar shape and provides a better basis as a surrogate for shape comparison.

Restricting the total number of non-hydrogen atoms was essential to the correspondence between MDDR and CSD. Our definition of shape is not size-independent, i.e. larger molecules have different shapes than smaller ones. In addition, larger molecules can adopt a greater variety of shapes than smaller ones and so the dimensionality of shape must increase. Future work will describe more closely the dependence on dimensionality and molecular size and its consequence for the fingerprint approach. Nonetheless, it was surprising that a reasonable window of heavy atom count for drug design (12−32 atoms) produced such a low number of reference shapes, leading to very practical storage and calculational requirements for shape-fingerprints.

The reference shapes are designed not only to cover shape-space but also to provide information about the population density of local regions of shape space. From the data in Figure 14 it can be seen that there is variation in the frequency that certain bits are turned on. When the fingerprint is generated, there is information on the 'popularity' of the shape, depending on whether it turns on common or infrequent bits. This can be used to anticipate the number of shape neighbors a molecule possesses. Such information is valuable in, for instance, designing screening sets. It might also prove useful in the analysis of affinity versus selectivity: an active molecule with a common shape might be less likely to be selective than one with a rarer shape.

One of the difficulties of complementing 2D methods by 3D techniques is the multiple-conformation problem. 2D approaches have unique representations; 3D has to deal with many. It has been shown that between 20 and 200 conformers are required to have a reasonable probability of representing the bioactive conformation of a drug-like molecule.[30,31] A common method with 2D fingerprints to speed processing is to 'fold' fingerprints. Each fingerprint is split into two, and the logical operation 'OR' is applied to each half. The reduced fingerprint has the characteristic that single bits represent multiple features. Shape-fingerprints could be treated in a similar way by folding all the fingerprints for a set of conformers into a single fingerprint. This, then, puts a 3D fingerprint on the same footing as a 2D fingerprint. It is unique to that molecule, not to a conformer of the molecule. One could also cluster conformation initially on shape and only store fingerprints of the cluster centers. It has not been lost on us that, once a conformational expansion is uniquely represented, shape-fingerprints provide a method of molecular 'encoding', i.e. storing an important aspect of a molecule that cannot be reverse engineered to its identity. Such molecular 'hash' functions have use in the safe exchange of chemical information.

The analysis of large data sets using a clear notion of molecular shape has not been previously possible because of the prohibitive computational expense of calculating shape similarity. The shape-fingerprint approach ameliorates this

problem. Work is underway on using these fingerprints in the interpretation of biological activity of molecules in terms of molecular shape. Retrospective work has shown that this has potential,[2] and it is anticipated that this will be a useful complement to existing 2D methods, which describe features of the chemistries of molecules. In particular, the utility in finding shape clusters will be examined. Most clustering methods require $O(N^2)$ operations and are the bottleneck for statistical analysis of large sets. Not only does the reduction of shape to a fingerprint greatly reduce the computational cost but also the metric nature of shape-space should be of utility, for instance in typical branch-and-bound methods.[32,33] Shape clusters will be ideal for a variety of methods aimed at profiling data sets, whether they are corporate collections, combinatorial libraries, or vendor compounds. Clustering could also be used to identify if certain classes of drugs possess a common shape or shapes.

A potential method for improving the precision of the fingerprint approach is to avoid the reduction of the explicit shape overlay Tanimoto value to binary form, i.e. by the storage of the exact distance from each reference shape to the shape being fingerprinted. This additional information might be expected to lead to a more precise mapping albeit at the expense of the computational efficiency of binary fingerprints. However, early work by the authors examined the use of such shape vectors and found numerical noise from the larger distances, i.e. dissimilar structures, tended to overwhelm the signal from similar reference structures. Filtering such information, for instance storing only distances below a threshold, might overcome these difficulties.

The method of constructing the set of reference shapes has proved fruitful. It has illustrated the dimensionality of the metric space and the universality of shapes chosen from a diverse set of molecules. As such, it is useful in itself as a fundamental, parameter and heuristic-free, measure of molecular diversity. However, the algorithm is not fast, requiring $O(N*M)$ operations where N is the total number of structures and M is the final number of reference structures and since M can be large, applying this clustering to very large data sets requires significant computation. However, we have shown than M is a very strong function of the Design-Tanimoto (Figures 1 and 3). As such, it takes much less time to construct a coarse representation of shape-space. There is nothing in the method that necessitates starting with a single structure in R, the set of representatives; one could just as easily start with a set generated with a coarser threshold. The advantage here is that many of the structures in S would remain assigned to the coarser representatives even with a higher Design-Tanimoto, greatly decreasing the number of comparisons. The understanding developed here should lead to highly efficient methods for the generation and evaluation of a library's diversity and coverage of shape-space.

We have not described the facility to include electrostatic information in the shape comparison of molecules. This can greatly improve the reliability of shape as an indicator of activity especially when used in conjunction with 2D measures.[2] Electrostatic comparisons also possess metric properties and are more amenable to partial-shape analysis (by defining local electrostatic domains). Currently, comparisons of potentials are performed after alignment by shape but could be done independently. The method described here could be duplicated to produce 'electrostatic fingerprints' that are independent of molecular shape, and such work will be presented elsewhere.

CONCLUSION

This work has shown that a binary fingerprint can encode the molecular shape of a compound. Consequently, only bit comparison operations are required to compute molecular shape similarity with little loss of accuracy. This is demonstrated in the ROC analysis of shape neighbor retrieval using shape-fingerprints against computation of explicit, optimal, volume overlap. The shape-fingerprint was designed by an analysis of the Cambridge Structural Database. This resulted in a set of reference shapes, and a small set of parameters that determine the optimal binary fingerprint. The transferability of these reference shapes and parameters was demonstrated by constructing fingerprints on a different database of molecules and obtaining similar results. This establishes that shape-fingerprints can be applied universally in computing shape-similarities of arbitrary, but size-constrained, molecules. The speed of determining molecular shape comparisons is now similar to standard 2D similarity measures and several orders of magnitude faster than typical 3D methods.

ACKNOWLEDGMENT

Dr. Brooke Magnanti is thanked for help in preparing this manuscript.

REFERENCES AND NOTES

(1) Rush, T. S.; Grant, J. A.; Nicholls, A. A New 3-D Scaffold Hopping Algorithm and its application to the ZipA-FtsZ protein−protein interaction. *J. Med. Chem.* Submitted for publication.
(2) Nicholls, A.; MacCuish, N. E.; MacCuish, J. D. Variable Selection and Model Validation of 2D and 3D Molecular Descriptors. *J. Comput.-Aided Mol. Des.* In press.
(3) Flower, D. R. On the properties of Bit String-Based Measures of Chemical Similarity. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 379−386.
(4) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(5) James, C. A.; Weininger, D.; Delany, J. Daylight Theory Manual 4.71, Daylight Chemical Information Systems, Inc., Irvine, CA, 2000.
(6) *UNITY*, version 4.4.2; Tripos Inc.: 1699 South Hanley Road, St. Louis, Missouri 63144, USA.
(7) McGregor, M. J.; Pallai, P. V. Clustering of Large Databases of Compounds: Using the MDL "Keys" as Structural Descriptors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443−448.
(8) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273−1280.
(9) Bradley, E. K.; Beroza, P.; Penzotti, J. E.; Grootenhuis, P. D. J.; Spellmeyer, D. C.; Miller, J. L. A rapid computational method for lead evolution: Description and application of alpha-adrenergic antagonists. *J. Med. Chem.* **2000**, *43*, 2770−2774.
(10) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. New 4-point pharmacophore method for molecular similarity and diversity applications: Overview over the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *J. Med. Chem.* **1999**, *42*, 3251−3264.
(11) McGregor, M. J.; Muskal, S. M. Pharmacophore fingerprinting. 1. Application to QSAR and focused library design. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 569−574.
(12) Grant, J. A.; Pickup, B. T. A Gaussian Description of Molecular Shape. *J. Phys. Chem.* **1995**, *99*, 3503−3510.
(13) Grant, J. A.; Gallardo, M. A.; Pickup, B. T. A Fast Method of Molecular Shape Comparison: A Simple Application of a Gaussian Description of Molecular Shape. *J. Comput. Chem.* **1996**, *14*, 1653−1666.
(14) *ROCS (Rapid Overlay of Chemical Structures)*, version 2.0; OpenEye Scientific Software: Santa Fe, New Mexico, U.S.A, 2004.

(15) Continuing work at AstraZeneca.
(16) Jones, H. D. Investigating the Metric Nature of Molecular Shape. Ph.D. University of Sheffield, 2004.
(17) Hanley, J. A.; McNeil, B. J. The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology* **1982**, *143*, 29−36.
(18) Bradley, A. P. The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms. *Pattern Recognit.* **1997**, *7*, 1145−1159.
(19) Jain, A. K.; Dubes, R. C. *Algorithms for Clustering Data;* Prentice Hall: 1988.
(20) Allen, F. H.; Davies, J. E.; Galloy, J. J.; Johnson, O.; Kennard, O.; Macrae, C. F.; Mitchell, E. M.; Mitchell, G. F.; Smith, J. M.; Watson, D. G. The Development of Versions 3 and 4 of the Cambridge Structural Database System. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 187−204.
(21) Allen, F. H.; Kennard, O. 3D search and research using the Cambridge Structural Database. *Chem. Des. Autom. News* **1993**, *8*, 31−37.
(22) *MDDR*, MDL Information Systems.
(23) *ClogP*, version 4.72; Daylight Chemical Information Systems Inc., Irvine, CA.
(24) Gasteiger, J.; Rudolph, C.; Sadowski, J. CORINA. 3-D atomic coordinates for organic molecules. *Tetrahedron Comput. Method.* **1992**, *3*, 537−547.
(25) *OMEGA*, OpenEye Scientific Software: Santa Fe, New Mexico, U.S.A.
(26) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of several chemical-structure file formats used by computer-programs developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, *32* (3), 244−255.
(27) Grant, J. A.; Pickup, B. T. Gaussian Shape Methods. In *Computer Simulation of Biomolecular Systems*; Van Gusteren, W., Weiner, P., Wilkinson, A. W., Eds.; Kluwer/Escom: 1998; pp 150−176.
(28) Mansfield, M. L.; Covell, D. G.; Johnson, O. A New Class of Molecular Shape Descriptors. 1. Theory and Properties. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 259−273.
(29) Haigh, J. A.; Grant, J. A.; Nicholls, A. Unpublished work.
(30) Boström, J. Reproducing the conformations of protein-bound ligands: a critical evaluation of several popular conformational searching tools. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 1137−1152.
(31) Boström, J.; Greenwood, J.; Gottfries, J. Assessing the performance of OMEGA with respect to retrieving bioactive conformations. *J. Mol. Graph. Mod.* **2003**, *21*, 449−462.
(32) Chávez, E.; Navarro, G.; Baeza-Yates, R.; Marroquín, J. L. Searching in metric spaces. *ACM Computing Surveys* **2001**, *33* (3), 273−321.
(33) Micó, L.; Oncina, J.; Carrasco, R. C. A fast branch and bound nearest neighbour classifier in metric spaces. *Pattern Recognit. Lett.* **1996**, *17* (7), 731−739.

CI049651V