J-CAMD 277

New molecular shape descriptors: Application in database screening

Andrew C. Good, Todd J.A. Ewing, Daniel A. Gschwend and Irwin D. Kuntz*

*School of Pharmacy, Department of Pharmaceutical Chemistry, University of California at San Francisco,
513 Parnassus Avenue, San Francisco, CA 94143-0446, U.S.A.*

Received 14 June 1994
Accepted 29 August 1994

Keywords: Geometric keys; Docking; Molecular similarity

Summary

Geometric descriptors are becoming popular tools for encoding molecular shape, for use in database screening and clustering calculations. They provide condensed representations of complex objects and, as a consequence, can usually be compared quite rapidly. Here we present a number of new descriptors and methods for the quantification of molecular shape similarity. The techniques are tested using two different biological systems, with particular emphasis on their potential utility as methods for prescreening shape-based database searches. Results are compared with data sets produced using the DOCK program. We find that such similarity evaluations are useful for finding molecules with complementary shape, and that they contain an enriched number of potential DOCK hits when compared to the original databases. Significant limitations in the utility of such DOCK prescreens are discussed, and potential solutions are considered.

Introduction

The lock-and-key approximation for receptor–ligand recognition forms an integral part of many computer-aided drug design applications. Central to this representation is the tenet that, for a ligand to interact with a given receptor, it must exhibit some degree of shape complementarity. The DOCK program developed by Kuntz et al. [1,2] uses the lock-and-key approach in docking rigid ligands into fixed receptor active sites. DOCK employs graph matching techniques [3] to fit 3D database structures into receptors of interest, utilizing active-site descriptions obtained from 3D receptor coordinate data. Molecular docking procedures of this form explore six degrees of freedom, allowing the examination of a large number of ligand–receptor orientations throughout the active site. However, as a consequence of the large number of possible geometries, most of this time is spent rejecting nonproductive (high-energy) binding modes. To combat this, methods are required which can be used to prescreen DOCK runs. Possibilities include employing chemical information during the search phase [4], using critical receptor features to focus the search (Desjarlais, R., private communication), and clustering the database in order to reduce the size of the search [5]. In this paper we explore the possibility of prescreening database searches based on shape similarity.

A number of shape descriptions have recently been developed to allow rapid similarity evaluations. Pepperell et al. [6] considered a number of similarity search techniques involving measures of interatomic distance. Fisanick et al. [7] developed descriptors which employ the geometric features of atom triangles, for example perimeter and area, in order to measure molecular shape. These descriptors (along with many others) were applied to searches of the Chemical Abstracts Service (CAS) registry substances. Bemis and Kuntz [5] calculated molecular shape, utilizing histograms of atom triangle perimeter data, and applied the resulting shape descriptions to database clustering. Of particular interest are the shape screens developed by Nilakantan et al. [8]. Their technique again uses the distribution of atom triplet distances within a molecule to describe molecular shape. Triangle side lengths are calculated for each atom triplet. These lengths are sorted and scaled by size and subsequently assigned a triplet value according to the equation

$$\mathrm{nt} = \mathrm{n1} + 1000\,\mathrm{n2} + 1\,000\,000\,\mathrm{n3} \qquad (1)$$

*To whom correspondence should be addressed.

where nt is the packed integer, and n1, n2 and n3 are the three digitized and sorted sides of the triplet. The triplet value is digitized and packed into a 2048 bit signature in order to save space. Signature similarity is determined using Eq. 2:

$$s = 2c / (na + nb) \tag{2}$$

s is the similarity, c is the number of bits in common, and na and nb are the number of bits in the template and database molecules. Triplets are recalculated on the fly for database molecules deemed similar enough to the template based on signature. A more accurate similarity value is then determined using the resulting triplet data and equations analogous to Eq. 2. Nilakantan et al. used their system both to prescreen DOCK searches and as a standalone shape search system. Norel et al. [9] have developed another variant of the triangle descriptor to calculate molecular shape. Pairs of atoms are chosen from a given structure, and triangles are constructed by adding a third vertex in the form of a molecular surface point. Triangle data for multiple atom pair–surface triangle descriptions are stored in hash tables, and used to rapidly screen for potentially complementary ligand–receptor interactions.

Here we explore ways in which triangles and other shape descriptors may be employed in order to undertake molecular similarity searches. These include atom and surface point triplets properties, angular information derived from surface point normals, and local curvature data. The descriptors are combined both in the form of histograms and bit strings. Searches are considered both from the standpoint of their ability to suggest new shape topologies and as tools for DOCK prescreens.

Materials and Methods

A number of molecular shape descriptions have been devised for use in similarity calculations. We first describe geometric measures based on molecular atom distributions. Then we outline descriptors which utilize molecular surface properties in order to measure shape. Note that for all techniques, molecular shapes are keyed using all (heavy and hydrogen) atoms. While this increased the time required to key molecular shape, the extra CPU cost was not prohibitive, and it was felt that the resulting description would be a more faithful structural representation. All calculations undertaken below were executed on an SGI Iris 4D-35.

Atom-based descriptions

Atom triangle bit strings (method 1(i))

This first shape description method bears some resemblance to the techniques applied by Nilakantan et al. Molecular shape is keyed through the calculation of triangle side lengths, determined from the consideration of all combinations of atom triplets within a given molecule. Atom–atom distances are digitized using a 1.0 Å range, and triplets containing distances greater than 30 Å are ignored. The sides of each passing triplet are sorted by length (long, intermediate and short), and the resulting values are used to turn on the variable in the logical array SHAPE(long,intermediate,short). Once all triplets have been calculated, the SHAPE array is packed into a bit string, which acts as the molecular shape description. The largest possible bit string is thus 30^3 (27 000) bits. Packing is structured, however, so that the number of bits required to store all 'on' bits for a given molecule is 30*30* (the highest occupied 'short' dimension of the SHAPE array). Around 4 h were required to key 9956 compounds of the Cambridge Structural Database (CSD) [10], converted to SYBYL MOL2 file format [11] with hydrogen atoms added. On average, just under a kilobyte of hard disk storage was required to store the bit string of each structure.

Atom triangle histograms (method 1(ii))

While the use of bit strings permits the storage of triangle side data, no information can be stored on the relative frequency with which each triangle occurs. In order to allow this, histograms of triangle shape data are required. It is not possible to store the lengths as for the atom triangle bit strings, since the resulting integer*2 array would require 15 times more storage space. A storage requirement of 15 kb per molecule is too large for a large database. To circumvent this problem, a new way to measure triangle shape has been devised. The technique employs a 2D shape array. The first dimension corresponds to the perimeter P of a given atom triplet triangle. The second dimension corresponds to the deviation of a given triangle from equality, utilizing the fact that, for a given perimeter P, the area of a triangle is given by

$$\text{Area} = (P/2\,(P/2 - \text{Side1})(P/2 - \text{Side2})(P/2 - \text{Side3}))^{1/2} \tag{3}$$

The maximum possible area for a given perimeter P is produced by an equilateral triangle

$$\text{Max Area} = (P^4/432)^{1/2} \tag{4}$$

The deviation of the area for a given triangle perimeter to the possible maximum area thus gives a measure of the triangle's deviation from equality, and hence an indirect measure of its shape. We stored information regarding the area as the ratio of the triangle area to the maximum possible area for a given perimeter, according to the equation

$$\text{Ratio} = \frac{\text{Area}}{\text{Max Area}} \exp - (((P/3 - \text{Side1})^2)^{1/2} + ((P/3 - \text{Side2})^2)^{1/2} + ((P/3 - \text{Side3})^2)^{1/2}/2P/3) \tag{5}$$

where the exponential term is used to differentiate triangles closer to equality. Without this correction, sensitivity is restricted to triangles approaching zero area. Many other triangle measures are possible, for example, summing the deviation of each side of a triangle from equality (one third of the total perimeter). The area technique was applied because it was considered the least ambiguous shape measure. Perimeters are binned in 1 Å ranges up to 50 Å in length. All longer perimeters are ignored. The area ratio is multiplied by 10 and digitized, producing 10 ratio bins. The resulting integer array is SHAPE-(ratio,perimeter). Each time triplet data are calculated, the associated array variable is increased by one. In this way a shape histogram of the molecule is built up. In order to save space, only occupied array integers are stored, i.e. up to and including the longest perimeter found. CSD key time was around 1.5 h, with just over half a kilobyte of storage space required per molecule.

Surface-based descriptions

The program SPHGEN forms part of the DOCK suite of programs [1]. SPHGEN creates sets of overlapping spheres within the receptor region of interest, in order to produce a negative image of the active site. DOCK matches sphere centers with molecular atom centers in order to superimpose potential ligands with the active site. The spheres generated by SPHGEN provide an active-site description which bears some similarity to the van der Waals (vdW) atom spheres used to represent molecular structures. However, the distributions of SPHGEN spheres and ligand atom spheres show little resemblance, since SPHGEN spheres tend to cluster near the walls of active sites, in regions of high curvature, while atom centers are more evenly distributed throughout a molecule. Direct comparisons of atom and sphere center triplets is thus ineffective, since the relative bin occupations of atom and sphere center molecular descriptions show significant divergence. A different approach is thus required which provides structural descriptions that are consistent for both SPHGEN and ligand atom spheres. Since SPHGEN spheres map the volume of the active site and vdW spheres map the volume of ligand molecules, the *surface features* of the two sphere types should provide the consistent description we are seeking. The method has the advantage of not being directly dependent on the number of atoms (or spheres) in the system, depending instead on surface shape. The resulting shape description is also versatile, since regions with no known ligand can be prescreened using ligand–active site comparisons, through the creation of pseudoligands based on clusters of SPHGEN spheres.

Surface generation

In order to map structural surface features, a method for placing uniform points on the molecular or active-site surface was required. The Connolly surface software, MS, was found to be most suitable for this purpose [12]. The program is able to take any sphere set and generate the required surface, which is critical for the SPHGEN sphere clusters. The probe radius can be adjusted so that only gross structural features are considered. Surface point density can be adjusted to allow easy optimization of histogram accuracy to generation time. Finally, the program source is readily available [13], allowing easy incorporation within any shape descriptor generation program.

Two major variables control Connolly surface calculations and need to be set in order to provide the best compromise between keying speed and shape descriptor accuracy:

(i) Probe size can have a significant effect on the number of dots used to describe a ligand. Small probes can create surfaces which include dips and bumps. The Connolly surface is discrete in nature, so subtle features such as these could produce significant differences in the number of surface points present on ligand or sphere clusters whose gross shapes are similar. To minimize this problem, a probe of 1.5 Å was chosen for surface creation.

(ii) In an ideal world, the choice of a high surface dot density would be automatic to ensure accurate shape description. However, because histogram calculation times increase rapidly as more points are added to the surface, high dot densities would lead to unacceptably long keying times. On the other hand, too low a density would lead to the possibility of incomplete shape descriptions. To determine the optimum surface dot density, multiple surfaces of varying density were generated for a number of different molecules. The distribution of inter-surface point distances was then measured, in order to determine the lowest dot density producing distance distributions similar to those of higher density surfaces. The results determined for the Brookhaven Protein Databank (PDB) [14] hydrogenated crystal structure of 4DFR-bound methotrexate [15] are shown in Fig. 1. These data are indicative of the behavior of the other molecules examined. A 1.00 dot/Å^2 density surface seems to be the lowest density producing results consistent with those of higher density. Surfaces with a density lower than 0.5 dots/Å^2 are unreliable (visual analysis suggests that the reentrant surfaces provide the major error source). Based on these data, we chose to make general use of the 1.00 dot/Å^2 density in the creation of surface shape descriptions.

Triangle construction

An alternative approach to triangle construction was designed for use with the surface data. Using a 1.00 dot/Å^2 dot density surface on methotrexate produces 326 surface points. If we were to undertake full triple calculations using such a system, we would need to calculate data for n(n–1)(n–2)/6 triangles. This converts to 5 721 300 triangles. Thus, the times required to key large databases would clearly be prohibitively long. In order to circumvent this problem, triples were calculated using the surface point centroid and all combinations of surface point pairs (centroid triangles). This has the advantage of requiring only n(n−1)/2 triangle calculations, or a total of 52 975 triangles for the methotrexate system, a saving of over two orders of magnitude.

Additional descriptors

In addition to employing centroid triangles, a number of additional measures were devised in the hope of improving shape description resolution.

(i) The directions of the surface normals provide implicit information about the relative orientations of the surface points making up part of any given centroid triangle. This directional information can be partially determined via calculations of the angle between the normals. Further definition can be obtained through calculation of the torsion angle between normals, relative to the chord linking the two surface points that make up the centroid triangle. An alternative method for calculating the relative normal alignment is to sum the normal vectors. The magnitude of the resulting vector imparts information regarding normal orientation.

(ii) In order to impart some information regarding the shape of surface local to any given point, techniques were devised for calculating a curvature value at any given surface point. This form of measure has been applied by Connolly in order to map the minimum and maximum shape functions (knobs and holes) across a protein surface [16]. Connolly used a grid calculation in order to evaluate the volume of a sphere centered at each surface point that is buried within the protein. This technique is too time consuming for use in database keying, so an alternative scheme has been devised. Curvature is measured by determining the coordinates of a point 1.5 Å along the surface point normal vector (henceforth known as the normal point). The angles between the surface point normal and the vectors from the parent surface point to all surface points within 3 Å of the normal point are then determined. The average angle is calculated together with the minimum and maximum angle. Curvature bounds were determined empirically through tests on a number of molecules. The bounds were varied until concave, convex and flat regions could be assigned in such a way as to look reasonable by visual inspection of an appropriately color-coded surface. These tests showed that a minimum angle of less than 65°, together with an average angle of less than 90°, worked well as a definition for concavity. An average angle of greater than 90° with a maximum angle of greater than 115° was used to define convex curvature. All other results were defined as flat. Saddle regions were considered to be flat, since there is a limit to the amount of shape definition that can be packed into a reasonable amount of storage space. Another local shape measure considered was a local shape point count. Counting the number of surface points within a set radius of a given normal point would give an indirect measure of local shape. The larger the number of points, the more concave the shape. This was not pursued, because of its sensitivity to different surface densities, but it is a potentially useful descriptor. There are many ways in which these measures can be used. By varying the angles and point count bounds associated with a given shape, the sensitivity of the calculation can be set according to specific user requirements.

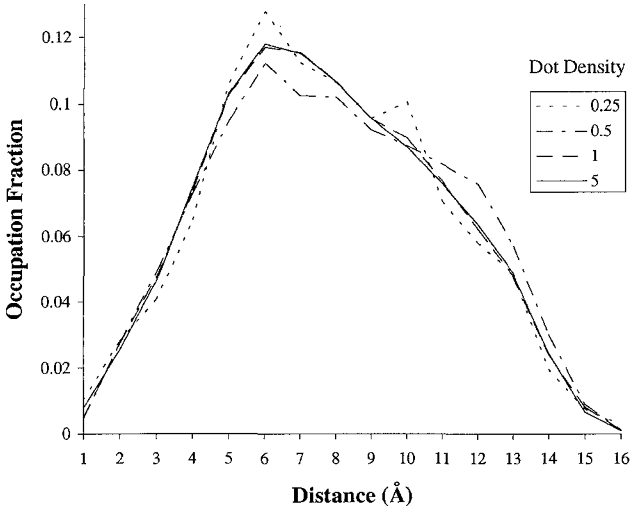

Fig. 1. Distribution of intersurface point distances for the MS surface description (1.5 Å probe) of methotrexate, using different dot densities.

Surface descriptor details

The shape descriptors described above have been combined in a number of ways to produce overall molecular shape descriptions.

Centroid triangle perimeter surface histograms (method 2(i)) This is the simplest histogram shape description, in which the perimeter data of the centroid triangles (see the section on triangle construction) are stored in 50 bins of 1 Å. CSD key time using a 1.00 dot/Å^2 dot density was around 3 h, with an average compound storage requirement of just over 0.1 kb.

Centroid triangle/normal angle surface histograms (method 2(ii)) This shape description employs the perimeter and area ratio descriptors described in method 1(ii). Centroid triangles (see above) are used. Perimeters are binned in 1.5 Å ranges up to 60 Å in length. The area ratio is multiplied by seven and digitized, producing seven area ratio bins. In addition, three angle bins (60°, 120°, 180°) for the angle between normals are included in the description. Using a 1.00 dot/Å^2 density, CSD key time was around 6 h, with each structure taking on average approximately 0.75 kb storage space.

Multi-surface descriptor bit string (method 2(iii)) This description contains all the shape measures described above. In addition to the centroid triangle perimeter and

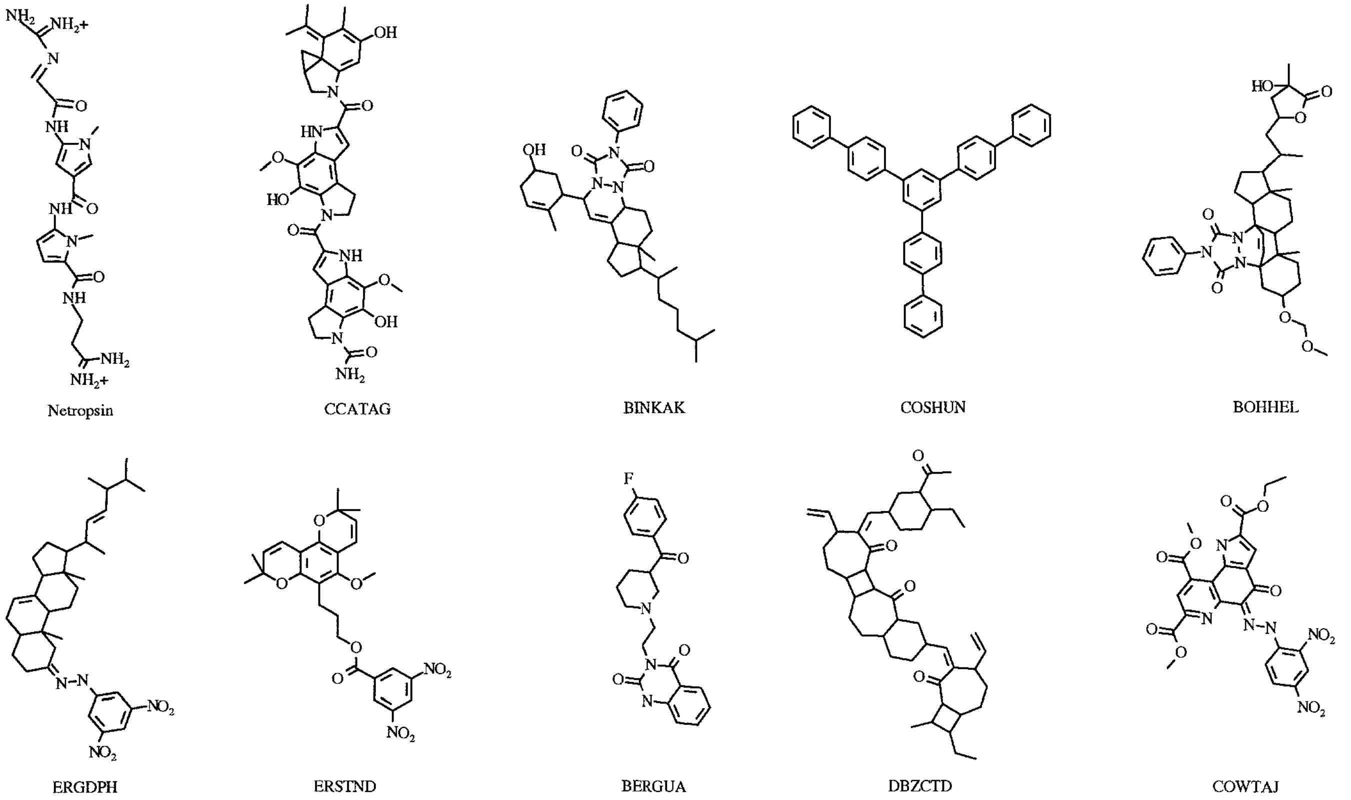

Fig. 2. Ten top hits from a DOCK search of the B-DNA minor groove.

area ratio, normal angle, normal torsion and curvature are included. The five-dimensional array is divided up as follows: 40 perimeter slots of 1.5 Å range, five area ratio slots, three 60° normal angle slots, six 60° normal torsion slots and six curvature slots. The curvature slots divide up according to the type of curvature present on the two surface points that make up a given centroid triplet, i.e. concave–concave / concave–flat / convex–convex, etc. The shape description is packed as a bit string according to the process described above. For a 1.00 dot/Å^2 density surface, CSD key time was around 12 h, with an average storage requirement of just under 1.5 kb. Once again the storage requirements make a histogram equivalent of the shape description untenable for large database keying.

Clustered triangle surface histograms (method 2(iv)) If we wish to use full (noncentroid) triangle descriptors in conjunction with high density surfaces, some form of surface point clustering is required. To achieve this, the following procedure has been devised. When the molecular surface is created, all surface points (contact, re-entrant and saddle) associated with a particular atom are clustered. All points within 1.5 Å of each other are assigned to the same cluster. All clusters with a surface area less than 0.5 Å^2 are removed. This cutoff is determined by dividing the number of cluster points by the surface dot density. The average coordinate positions are then determined for all remaining clusters, and these are stored together with the number of points associated with the cluster. Most atoms will produce only one cluster. Some atoms with two surface faces, such as benzene ring carbons, will produce two. Triangles are then determined between clustered points using the same general procedure as in method 1(ii). A 30-slot perimeter of 2.0 Å range is used, and arrays are augmented by adding the multiple of the number of surface points associated with the points making up the triangle under consideration. For example, a triangle of cluster points with 10, 12 and 17 associated points would result in a bin augmentation of 10*12*17. In this way, triangles are weighted by the amount of surface area associated with them. The number of points associated with a molecule tends to be similar to the number of atoms and, as a consequence, high surface densities can be used during keying. Since the clusters are not projected back onto the surface, the resulting points will tend to be slightly offset. Since we are comparing ligand with ligand, however, this was not considered to be a problem, since similar molecules should exhibit similar offsets. CSD keying using 6.00 dots/Å^2 surface density required 8 h, with around 0.3 kb storage space used per compound.

Similarity calculations

The following equations were applied to similarity evaluations:

$$S_{AB} = \sum_{i=1}^{n} B_i^2 \Big/ \sum_{i=1}^{m} \max(T_{A_i}, T_{B_i})^2 \qquad (6)$$

$$S_{AB} = \sum_{i=1}^{n} B_i^2 \Big/ \sum_{i=1}^{l} T_{A_i}^2 \qquad (7)$$

B equals the shape description overlap between equivalent histogram bins, or the number of matching bits for equivalent packed integers. T_A and T_B are the values for the individual descriptions for the same histogram bins / packed integers. The variables l, m and n are the highest occupied bin / packed integer of the template description, larger description and smaller description, respectively. Equation 6 uses as its denominator the maximum of two T values and is applied as a measure of overall similarity. This equation is derived from a similar formula, used to measure 3D molecular similarity of electrostatic potential [17]. Equation 7 uses the template description properties as its denominator, and is similar to one of the formulae utilized by Nilakantan et al. [8].

Bit string similarity evaluations

Bit string similarity calculations have been undertaken using the following procedure. An array (NUMON) was precalculated, containing the number of binary bits turned on for a given integer value. When two shape descriptions are compared, a bitwise AND is applied to each equivalent pair of 16 bit integers within the bit string. The resulting value is used to call the NUMON array, and the resulting array value is used as the overlap value for the integer comparison. Integer pairs are called in order until all have been compared. For the bit strings, Eqs. 6 and 7 are modified through the multiplication of numerator and denominator by the square of the current integer being compared. For example, if the third bit string integers are being compared, numerator and denominator overlap values are multiplied by 9. In this way, bits describing larger perimeter triplets with longer shortest sides are more heavily weighted. This is desirable, since these triangles tend to contain more shape information.

Histogram similarity evaluations

Overlap values for histogram description are determined by simply adding the minimum of the two description values in a given bin to the numerator term of the similarity equation. An additional function was also added for histogram comparisons in the form of bin normalization. Histogram bin values are normalized by dividing the bin value by the total value of the occupied bins. In doing this, sensitivity to absolute differences in bin occupation is reduced, while information regarding the distribution of descriptors is retained.

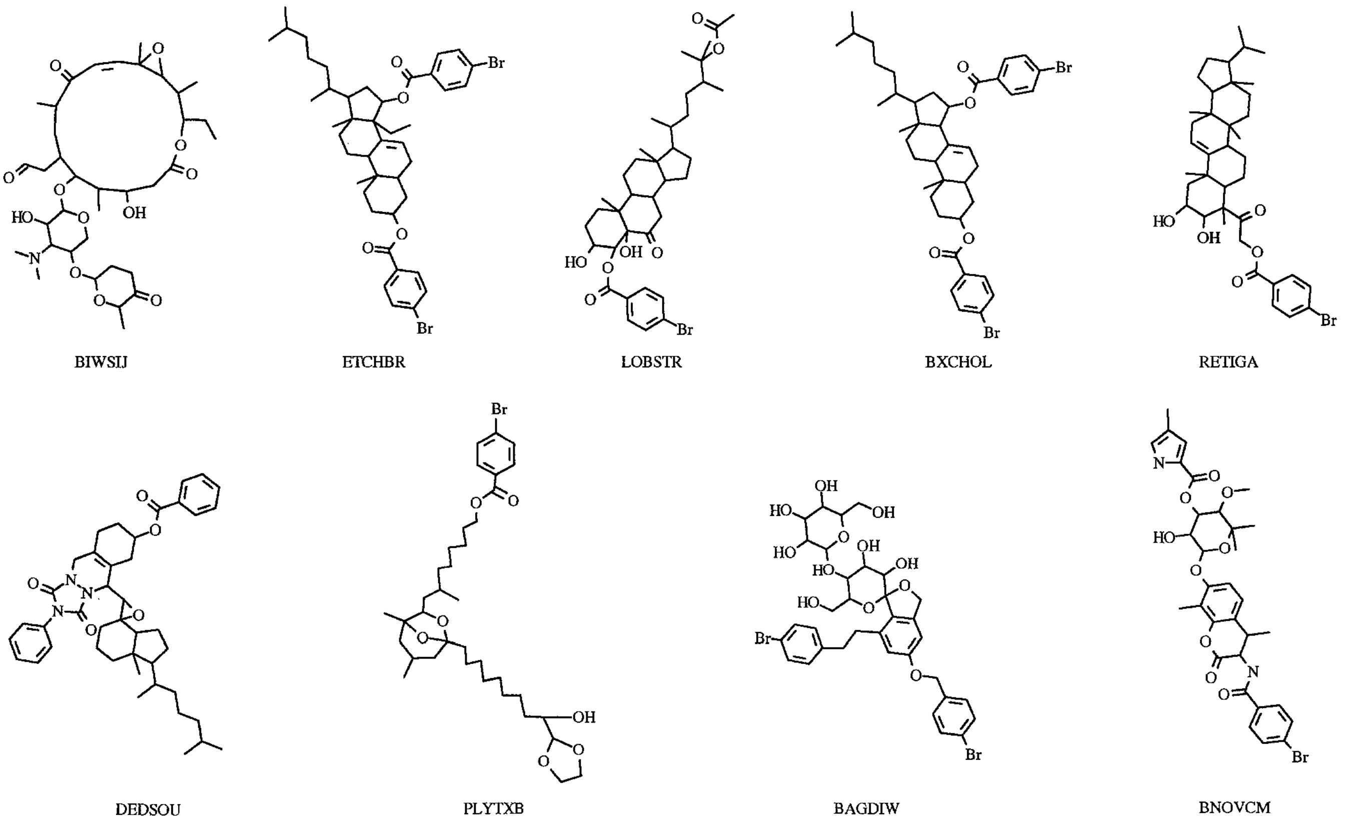

Fig. 3. Nine of the ten top hits obtained by the netropsin template search undertaken by Nilakantan et al. The missing hit is CCATAG (see Fig. 2).

Descriptor tests

In order to test the geometric descriptors described above, two investigations were undertaken. The first study repeated the experiment used by Nilakantan et al. [8] to test the utility of their shape descriptors. The DNA minor-groove ligand netropsin was used as the template to search the CSD. The results obtained from these searches were compared with those obtained in a previous DOCK search of the B-DNA minor groove, and also with the findings obtained by Nilakantan et al. In the second investigation, prescreen searches were undertaken using methotrexate on a section of the Fine Chemicals Database (FCD) [18], and the results were compared with those obtained from a DOCK search of the methotrexate binding region of 4DFR.

Netropsin study

In this study we used netropsin as the template molecule to search the CSD for molecules of similar shape. The PDB structure of netropsin bound to the minor groove of B-DNA (6BNA) [19] was used as the basis for the template. Hydrogens were added to the molecule and the resultant structure was minimized using the SYBYL molecular modelling program [11]. This was done both to ensure a low-energy structure, and to guarantee differences with the CSD version of the netropsin molecule. Shape descriptions were then calculated using this structure for all the geometric descriptors, and the resulting templates were used to search the 9956 structures of our CSD. Comparisons were made between the structure rankings obtained from these searches and the top 10 hits obtained in a previous DOCK search of the B-DNA minor groove [20] (Fig. 2). The results were also compared with the top hits obtained by Nilakantan et al. in their equivalent investigation (Fig. 3). The findings obtained utilizing Eq. 6 are shown in Table 1, normalized histogram results using Eq. 6 are listed in Table 2, and data retrieved applying Eq. 7 are shown in Table 3.

Methotrexate study

For this investigation methotrexate was used as the template molecule to search a 7302 compound section of the FCD. The conformer of methotrexate (with hydrogens added) bound to 4DFR [15] was used as the basis for the template. As before, shape descriptions were calculated using this structure for all the geometric descriptors, which were then used to search the FCD segment. A DOCK run was then undertaken on 4DFR, using a contact score grid with a 2 Å extent surrounding the bound methotrexate molecule, searching the same section of the FCD. The heavy atoms of the bound methotrexate structure were used as the receptor site points. The top 300 hits were retained. In addition to these calculations, a SPHGEN run was undertaken on the 4DFR active site using a 1 dot/Å^2 surface. Spheres whose centers were found within the vdW surface of the bound methotrexate molecule were retained in order to create a pseudoligand.

TABLE 1
RANKING OF TOP CSD DOCK (MOLECULES 1–10)/SCALED ATOM TRIPLE[a] SCREEN HITS, USING OVERALL SIMILARITY SEARCH (EQ. 6) AGAINST THE NETROPSIN TEMPLATE

	Technique (see Methods)					
Molecule	1(i)	1(ii)	2(i)	2(ii)	2(iii)	2(iv)
NETRSN	29	2	17	1	26	1
CCATAG	338	1630	2915	1827	297	2435
BINKAK	20	1751	408	248	41	264
COSHUN	541	185	1612	2171	742	240
BOHHEL	24	2561	455	236	55	705
ERGDPH	227	2224	410	791	323	414
ERSTND	359	330	533	278	429	249
BERGUA	622	107	223	159	809	677
DBZCTD	13	2870	1723	1098	37	2866
COWTAJ	183	124	552	381	289	325
DOCK hits in top 5% of prescreen	8	5	5	6	8	6
BIWSIJ	5	3067	2690	2109	25	3897
ETCHBR	356	2941	2908	2067	705	3561
BXCHOL	692	2735	2550	1593	924	3218
RETIGA	4	2611	847	415	7	318
DEDSOU	281	2893	2513	1900	260	3135
PLYTXB	26	2833	1075	477	16	2692
BAGDIW	21	1458	3077	2203	173	2461
BNOVCM	89	281	2956	2028	238	326

The number of DOCK hits present in the top 5% of the prescreen is also given. Molecule LOBSTR, ranked fourth by scaled atom triples to netropsin, is not present in our version of the CSD, and has thus been excluded from the results.

[a] Molecules 11–18 of Nilakantan et al. (Ref. 8).

TABLE 2
RANKING OF TOP CSD DOCK (MOLECULES 1–10)/SCALED ATOM TRIPLE[a] SCREEN HITS, USING NORMALIZED HISTOGRAM SEARCH AGAINST THE NETROPSIN TEMPLATE WITH EQ. 6

	Technique (see Methods)			
Molecule	1(ii)	2(i)	2(ii)	2(iv)
NETRSN	2	94	1	1
CCATAG	219	329	38	454
BINKAK	47	68	96	160
COSHUN	79	83	689	78
BOHHEL	121	62	42	164
ERGDPH	144	348	572	498
ERSTND	590	809	419	495
BERGUA	308	627	456	200
DBZCTD	52	40	83	226
COWTAJ	442	530	272	600
DOCK hits in top 5% of prescreen	9	8	8	9
BIWSIJ	60	34	121	321
ETCHBR	227	431	195	521
BXCHOL	122	546	208	400
RETIGA	117	36	56	178
DEDSOU	164	67	257	421
PLYTXB	33	76	12	146
BAGDIW	238	5	53	291
BNOVCM	38	161	18	212

The number of DOCK hits present in the top 5% of the prescreen is also given. Molecule LOBSTR, ranked fourth by scaled atom triples to netropsin, is not present in our version of the CSD, and has thus been excluded from the results.

[a] Molecules 11–18 of Nilakantan et al. (Ref. 8).

These spheres were used as the basis for a template which was keyed using method 2(ii). The resulting shape description was also used to search the FCD segment. The positions of the DOCK hits within each shape search list were then determined. This was done in order to test prescreen potential, through calculation of the degree of enrichment of the shape search lists, compared to random extraction of molecules from the database. The results obtained using Eq. 7 are shown in Fig. 4. Data for Eq. 6 using normalized histogram descriptions are shown in Fig. 5. The percentage of the top 50 DOCK hits found in the top 10% of the prescreen hit list for each technique are listed in Table 4.

Discussion

The calculations detailed above illustrate the potential utility of the geometric descriptor-based shape searches. From a DOCK prescreen perspective, when utilizing Eq. 6 with normalized histograms (Table 2) or applying Eq. 7

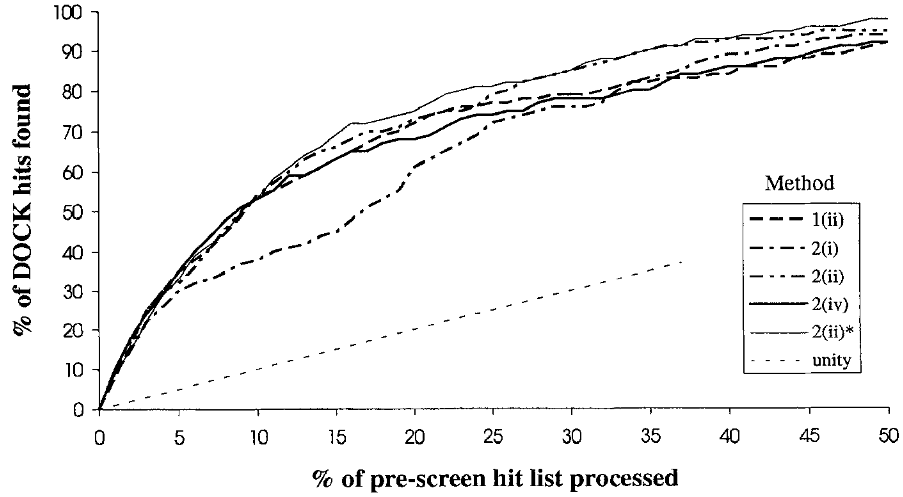

Fig. 4. Plot of the percentage of DOCK hits found in a given top percentile rank of the prescreen search lists, using a normalized histogram search against the methotrexate template, with Eq. 6. Method 2(ii)* indicates use of an SPHGEN sphere pseudoligand as template.

TABLE 3
RANKING OF TOP CSD DOCK (MOLECULES 1–10)/SCALED ATOM TRIPLE[a] SCREEN HITS, USING SUBSTRUCTURE SEARCH (EQ. 7) AGAINST THE NETROPSIN TEMPLATE

	Technique (see Methods)					
Molecule	1(i)	1(ii)	2(i)	2(ii)	2(iii)	2(iv)
NETRSN	83	364	229	106	159	258
CCATAG	1 (joint)	26	17	20	10	16
BINKAK	28	21	65	100	101	44
COSHUN	121	121	62	237	495	84
BOHHEL	40	27	123	69	134	34
ERGDPH	188	172	124	455	153	151
ERSTND	394	610	807	568	543	373
BERGUA	766	1326	1170	1097	952	2055
DBZCTD	46	20	83	27	57	25
COWTAJ	195	807	429	296	326	477
DOCK hits in top 5% of prescreen	9	8	8	8	8	9
BIWSIJ	38	22	99	41	45	14
ETCHBR	12	2	137	46	6	121
BXCHOL	7	14	114	56	30	9
RETIGA	29	10	86	34	39	67
DEDSOU	6	6	112	44	35	17
PLYTXB	37	13	11	12	24	18
BAGDIW	18	9	64	5	27	6
BNOVCM	33	61	12	26	21	57

The number of DOCK hits present in the top 5% of the prescreen is also given. Molecule LOBSTR, ranked fourth by scaled atom triples to netropsin, is not present in our version of the CSD, and has thus been excluded from the results.

[a] Molecules 11–18 of Nilakantan et al. (Ref. 8).

(Table 3), at least eight of the top 10 DOCK hits are found in the top 5% of the prescreen hit list. In most cases the remaining one or two structures appear in the top 10% of the hit list. The use of Eq. 6 with unmodified histogram descriptions leads to less satisfying results (Table 1). This is to be expected, however, since only structures that are similar both in *shape* and *size* will register as similar (hence the high ranking of the CSD netropsin molecule in Table 1). While this is useful for certain search criteria, for example if one were searching for more exact shape matches of netropsin, it is of limited utility for DOCK prescreens, where exact size matches are not prerequisite. Similarly, the DOCK methotrexate prescreen comparisons illustrated in Figs. 4 and 5 show a significant data set enrichment, with around 50% of the top DOCK hits generally appearing in the top 10% of the prescreen hit lists. The enrichment is even greater when the top 50 DOCK hits are considered in isolation, with generally over 60% and up to 80% of the hits located in the top 10% of the prescreen

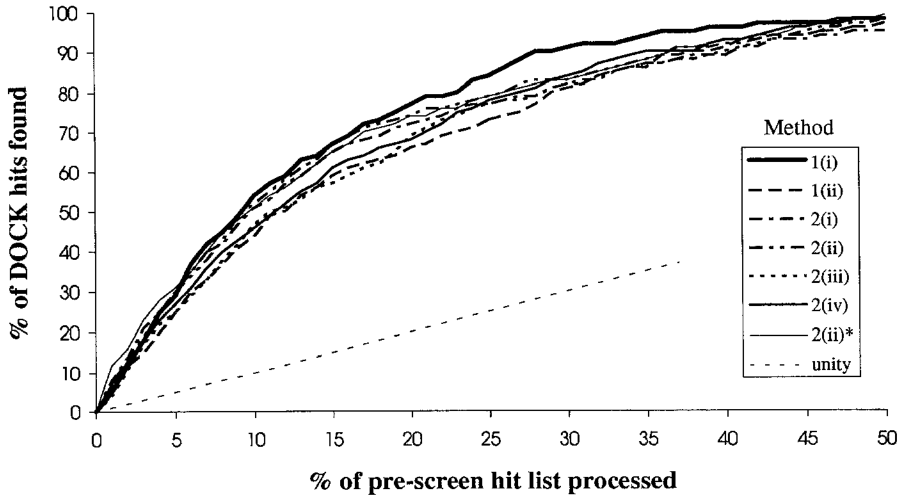

Fig. 5. Plot of the percentage of DOCK hits found in a given top percentile rank of the prescreen search lists against the methotrexate template, using Eq. 7. Method 2(ii)* indicates use of an SPHGEN sphere pseudoligand as template.

TABLE 4
PERCENTAGE OF TOP 50 DOCK HITS FOUND IN THE TOP 10 PERCENT OF THE PRESCREEN SEARCH LISTS FOR THE METHOTREXATE TEMPLATE FCD SEARCHES

Technique	% of top 50 DOCK hits in top 10% of prescreen	Technique	% of top 50 DOCK hits in top 10% of prescreen	Technique	% of top 50 DOCK hits in top 10% of prescreen
1(ii)[a]	60	2(ii)[b]	68	2(ii)[c]	74
2(i)[a]	48	1(i)[c]	80	2(iii)[c]	62
2(ii)[a]	66	1(ii)[c]	64	2(iv)[c]	70
2(iv)[a]	62	2(i)[c]	68	2(ii)[d]	66

[a] Used in conjunction with Eq. 6 and normalized histograms.

[b] Used in conjunction with Eq. 6 and normalized histograms. An SPHGEN sphere pseudoligand was used as the template.

[c] Used in conjunction with Eq. 7.

[d] Used in conjunction with Eq. 7. An SPHGEN sphere pseudoligand was used as the template.

rankings. These figures do not, however, represent a 5–8-fold enrichment, since the smallest maximum ligand atom pair distance to maximum receptor site point pair distance ratio was set to 0.8, to ensure that only molecules capable of spanning the methotrexate binding site were docked. This constraint removed around 40% of the database, thus the effective enrichment is around 3–5-fold. DOCK hits that appear lower in the list tend to be structures either significantly larger than methotrexate or structures that have been assigned a divergent binding mode. The results obtained using the SPHGEN sphere-derived templates compare well with those obtained using the explicit methotrexate structure (see Figs. 4 and 5, method 2(ii)*), illustrating the potential of the technique for probing previously unexplored active-site regions. The data are similar to the results obtained by Nilakantan et al. in their equivalent search using HIV-protease with the inhibitor MVT-101. Their enrichment is a little higher, but this is to be expected since MVT-101 is a larger inhibitor, and hence a more discriminating shape, and HIV-protease has a more enclosed active site, allowing fewer alternative binding modes. The use of these prescreens thus offers a useful way to focus DOCK searches onto particular regions of the site, while also significantly reducing the size of the database to be searched.

A comparison of results between shape measures suggests little difference in descriptor quality. The simplest method, 2(i), which only uses triplet perimeter data, appears to produce results which are a little less discriminating than the other techniques. This can be seen by its inferior prescreen performance (Fig. 4), its lesser ability to discriminate netropsin from the rest of the CSD (Table 2), and the fact that only 48% of the top 50 DOCK hits are found in the top 10% of the 2(i)-normalized histogram prescreen (Table 4). In general, however, all techniques

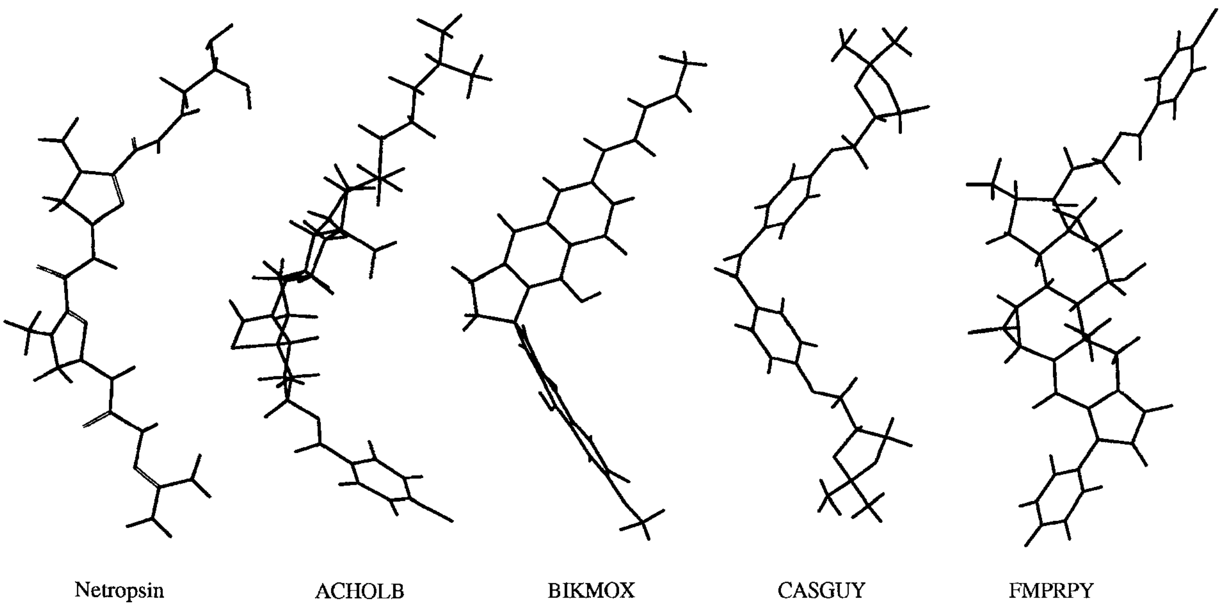

Fig. 6. Four structures from the top 50 hits of Eq. 6, i.e., a normalized histogram search against the methotrexate template, method 2(ii), exhibiting shape similarity to netropsin.

behave similarly, although if one had to pick a winner, method 1(i) would edge ahead because of its high enrichment performance in locating the top 50 DOCK hits of the methotrexate search (Table 4). In the end, however, descriptor choice is best left to user preference and available hard disk space.

One of the major strengths of these searches is their speed. Searches using these geometric descriptors proceed at between 500–2000 ligands per second, thus results can be obtained with minimal CPU cost. The hit lists also have other uses besides their application in DOCK prescreens. Nilakantan et al. showed how four of their top 10 hits could be minimized to fit to the B-DNA minor groove (ETCHBR, RETIGA, BIWSIJ and BNOVCM – see Fig. 3). These structures are generally found to appear in the top 50 of the substructure searches undertaken against netropsin (Table 3). Figure 6 illustrates four other top 50 ranking structures from the 2(ii)-normalized histogram search, with significant shape similarity to netropsin. These shape searches thus offer a fast way to determine possible shape templates that DOCK may miss because of a few bad contacts that could be alleviated through minimization.

Weaknesses and ways forward

While we have shown the potential utility of the geometric descriptors as prescreens, it is important to understand their limitations.

The fact that these descriptions do not contain any chemical information is often cited as a potential weakness of the geometric descriptors. To explore this point, all ligand points or receptor surface points considered to be involved in the binding mode of interest were coded according to the polarity of the atom to which the point was associated: positive, negative and hydrophobic (hydrophobic points are all nonpolar atom-associated points > 1.5 Å from the nearest polar surface point). These codes were then used to create six additional histograms / bit strings: positive–positive points only, positive–hydrophobic points only, etc. In this way a description of property shape could be built up in conjunction with overall shape. Similarity could be calculated using Eq. 7 for each equivalent property description, and an average property similarity determined. The technique was tested with a search of the FCD on the active site of the PDB structure 2GBP [21], which is small and highly polar. The top hits included isomers of glucose, fructose and other small sugars that are very similar to the glucose substrate. While initially these results were satisfying, the weaknesses of the technique are readily apparent. Because six further descriptors are required for each structure, storage requirements for the data can become a significant consideration. Also, the property description is only valid for the conformation keyed. As a consequence, this form of property measure is far less elegant than the flexible property descriptions used in 3D pharmacophore searches [22–24], where flexibility can be accounted for on the fly. Thus, it will likely prove most efficient to use geometric measures of shape in conjunction with pharmacophore-based searches, when chemical information is to be used as a constraint in the search.

The descriptors were shown to perform well in placing high-scoring DOCK hits high in the prescreen rankings. The netropsin B-DNA system is the perfect system for such a prescreen approach, since one is simply attempting to fit structures to the DNA minor groove. The number of alternative binding modes is thus severely limited, making netropsin an excellent prescreen template. The methotrexate system, however, is far from perfect. While we have shown that it is possible to focus on a particular binding mode with some success, there are many other potential binding modes possible within the 4DFR active site. Some attempt was made to divide the site up into pockets and link the pockets into a series of potential binding orientations. For the 4DFR system this yielded upwards of a 100 potential templates. If one assumes that 5% of each template hit list is retained, it is obvious that we soon end up making virtually no savings at all! Therefore, as the prescreens currently stand, an idea of the basic binding mode or a small enclosed binding site is required if one is to be successful.

To be a more generally useful tool within DOCK, the descriptors need to be meshed more tightly with the matching process of the program. The techniques developed by Norel et al. [9] provide an elegant method through which this may be achieved, using their shape measures as a technique for estimating match quality. The methodology has been shown to speed up the docking process, but requires a large amount of storage space to describe ligand shape. This approach is being combined with surfaces made up of sparse critical points developed by Lin et al. [25], in order to create a more efficient shape representation (Nussinov, R., private communication). Nevertheless, storing large databases of shape descriptions would require a large investment in hard disk space. We are currently investigating a variant of this approach, in which the shape local to each ligand atom and receptor site point are compared. Only atoms and site points with similar local shape environments will then be allowed to go forward to the DOCK matching process. In this way the geometric descriptors can be applied to any potential ligand–receptor orientation, the system essentially being used in order to direct the docking process. The storage requirements are lessened, since atoms are used as the descriptor anchors, rather than atom pairs. It is clear that further study is required in order to provide a shape description that is discriminating, requires little storage space, and yet has general applicability within the docking process.

Conclusions

We have developed a series of geometric descriptors for application both as DOCK prescreens and as standalone techniques for determining molecules of similar shape. These techniques are able to offer 3–5-fold enrichments as DOCK prescreens, but more work is required to enhance their general applicability.

Acknowledgements

Thanks go to Dr. Guy Bemis for help in the early stages of this work, and to Tripos Associates for making the Sybyl modelling package available to us. This work has been funded through NIH Grants GM31497 and GM39552 (G.L. Kenyon, P.I.).

References

1 Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R. and Ferrin, T.E., J. Mol. Biol., 161 (1982) 269.

2 Desjarlais, R.L., Sheridan, R.P., Seibel, G.L., Dixon, J.S., Kuntz, I.D. and Venkataraghavan, R., J. Med. Chem., 31 (1988) 722.

3 Shoichet, B.K., Bodian, D.L. and Kuntz, I.D., J. Comput. Chem., 13 (1992) 380.

4 Shoichet, B.K. and Kuntz, I.D., Protein Eng., 6 (1993) 723.

5 Bemis, G.W. and Kuntz, I.D., J. Comput.-Aided Mol. Design, 6 (1992) 607.

6 Pepperrell, C.A. and Willett, P., J. Comput.-Aided Mol. Design, 5 (1991) 455.

7 Fisanick, W., Cross, K.P. and Rusinko, A., J. Chem. Inf. Comput. Sci., 6 (1992) 664.

8 Nilakantan, R., Bauman, N. and Venkataraghavan, R., J. Chem. Inf. Comput. Sci., 33 (1993) 79.

9 Norel, R., Fischer, D., Wolfson, H.J. and Nussinov, R., Protein Eng., 7 (1994) 39.

10 Allen, F.H., Bellard, S., Brice, M.D., Cartwright, B.A., Doubleday, A., Higgs, H., Hummelink-Peters, B.G., Kennard, O., Motherwell, W.D.S., Rodgers, J.R. and Watson, D.G., Acta Crystallogr., B35 (1979) 2331.

11 SYBYL, Version 6.0, Tripos Associates, St. Louis, MO, 1994.

12 Connolly, M.L., J. Appl. Crystallogr., 16 (1983) 548.

13 Quantum Chemistry Program Exchange, Program No. 14291, Bloomington, IL.

14 Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer Jr., E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasomi, M., J. Mol. Biol., 112 (1977) 535.

15 Bolin, T.J., Filman, D.J., Matthews, D.A., Hamlin, R.C. and Kraut, J., J. Biol. Chem., 257 (1982) 13650.

16 Connolly, M.L., Biopolymers, 25 (1986) 1229.

17 Petke, J.D., J. Comput. Chem., 14 (1993) 928.

18 FCD-3D. Available from Molecular Design Ltd, San Leandro, CA.

19 Kopka, M.L., Yoon, C., Goodsell, D., Pjura, P. and Dickerson, R.E., Proc. Natl. Acad. Sci. USA, 82 (1985) 1376.

20 Grootenhuis, P.D.J., Kollman, P.A., Seibel, G.L. and Desjarlais, R.L., Anti-Cancer Drug Des., 5 (1990) 237.

21 Vyas, N.K., Vyas, M.N. and Quiocho, F.A., Science, 242 (1988) 1290.

22 Murrall, N.W. and Davies, E.K., J. Chem. Inf. Comput. Sci., 30 (1990) 312.

23 Hurst, T., J. Chem. Inf. Comput. Sci., 34 (1994) 190.

24 Moock, T.T., Henry, D.R., Ozkabak, A.G. and Alamgir, M., J. Chem. Inf. Comput. Sci., 34 (1994) 184.

25 Lin, S.L., Nussinov, R., Fischer, D. and Wolfson, H.J., Proteins, 18 (1994) 94.