Database Searching on the Basis of Three-Dimensional Molecular Similarity Using the SPERM Program

Nicholas C. Perry and Vincent J. van Geerestein*

Department of Computational Medicinal Chemistry, Organon International B.V., AKZO Pharma Division, P.O. Box 20, 5340 BH Oss, The Netherlands

Received July 30, 1992

An efficient method of determining the maximum in a similarity function between two molecules with respect to their mutual orientation in 3D rotational space is described. The similarity is calculated by comparison of gnomonic projections of molecular property values onto the vertices of a tessellated icosahedron. The optimal number of vertices required to obtain a reliable property profile is determined. A detailed description is given of a nondegenerate method of scanning rotational space by the use of a polar grid and application of icosahedral symmetry. The method is implemented in the SPERM program, which allows for searching 3D structural databases for molecules showing shape similarity to a given target structure. Search times may be reduced by the selection of suitable screen properties. The results of SPERM database searches using tetrodotoxin and netropsin as target structures are discussed.

1. INTRODUCTION

The concept of molecular similarity is an important one for rational drug design in pharmaceutical research.[1] In the search for new lead compounds, we might suppose that potential candidates would be similar in some way to known active compounds. This has led to a desire to search large databases of three-dimensional (3D) structures and rank molecules according to their similarity to a known target structure, which is typically a small organic molecule with a known pharmacological profile.[2] This requires a method of determining molecular similarity on a time scale which is appropriate for application to databases of several hundred thousand structures.

The 3D molecular similarity is calculated as a function of molecular properties that depend on the arrangement of the atoms in 3D space.[3] Here we describe a fast and efficient approach for optimizing the 3D similarity in Euclidean space between randomly oriented structures. Most such approaches employ the same fundamental procedure.[4] For each of the molecules to be compared, a property is calculated at each of a set of points in 3D space. By comparison of these two sets of values, a similarity index is calculated, and the orientation of one of the molecules is then optimized to give a maximum in the similarity function. Differences between the approaches arise from the following:

The molecular property used in the comparison
The method of sampling the property in 3D space
The method of calculating the similarity index
The method of sampling and optimizing the orientation in 3D space

We have previously described the SPERM program (from Superposition by **PERM**utation) in which the criteria for selecting these options are those appropriate for the efficient searching of 3D molecular databases.[5] Here we explain the methodology in more detail and describe a number of significant improvements that have been made to the original program.

2. METHODOLOGY

2.1. Choice of Molecular Property. The molecular property is typically a function of the 3D atomic coordinates and parameters associated with the atom and/or bond types. It is generally chosen to represent either the 3D steric, electrostatic, or lipophilic profile of the molecule, though in principle it may be any calculable property. Two main factors are involved in determining the choice of molecular property to be used in the similarity calculation for a database search:

The objective of the comparison
The speed with which the property can be calculated

Our ultimate objective is to discover new molecules which will bind effectively at the same receptor site as the target molecule. Their binding affinity will depend on a number of factors, including steric, electrostatic, and hydrogen-bonding characteristics. This suggests that the optimum choice should be some weighted combination of the appropriate factors. The disadvantage with this approach is that many molecules of potential interest will show a low measured similarity to the target molecule even though a minor chemical modification might dramatically increase the similarity. For this reason, we concentrate on measures of the steric profile on the assumption that many of our 'hits' may be chemically modified to improve the overall profile without significantly changing the steric properties of the molecule. Such properties require only the calculation of distances and may be quickly computed. They are therefore highly suitable for use in database searching.

2.2. Sampling of Molecular Property in 3D Space. The molecular property must be sampled at a number of points on a grid in 3D Euclidean space. Many similarity methods employ a Cartesian grid, but the number of points required to gain sufficient resolution in order to locate the global optimum similarity is prohibitive for database applications.[6] Chau and Dean[7] suggested an alternative grid comprising points on the surface of a sphere, the center of which coincides with the center (geometric or mass) of the molecule. In this case, the property at some location along a ray from the center of the sphere to a surface point (e.g., where it cuts the van der Waals surface) is associated with the grid point. This technique is known as gnomonic projection. In our applications we generally use the minimum distance (SURDIS) from the point to the molecular surface (VS_1 in Figure 1) or the radial distance (RADDIS) from the point to the surface (VS_2).

0095-2338/92/1632-0607$03.00/0

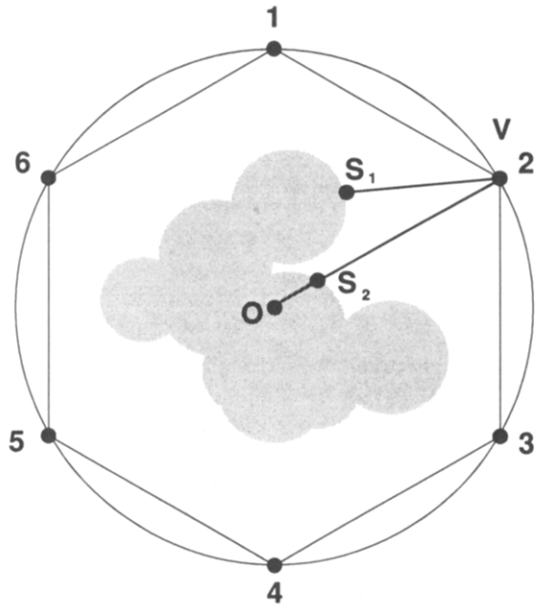

Figure 1. 2D analogy of gnomonic projection of molecular properties. The minimum distance from the vertex to the molecular surface (SURDIS) is given by VS_1, and the radial distance (RADDIS) is VS_2.

It is important that the points are (approximately) evenly distributed over the surface of the sphere and that the number is sufficient to ensure that the property profile obtained is essentially independent of the orientation of the molecule with respect to the points. The highest-order regular polyhedron is the dodecahedron, and consequently it is impossible to have more than 20 perfectly evenly distributed surface points. For reasons described in the following section, it is important that the points are related by icosahedral symmetry, and thus we use a tessellated icosahedron to provide additional surface points. These are generated by extending the vectors from the center of the sphere through the points generated by dividing the triangular icosahedral faces into smaller triangles as shown in Figure 2. The tessellation frequency is defined as the number of segments into which each icosahedral edge is divided and the number of vertices (n) is related to the tessellation frequency (f) by the formula:

$$n = 2 + 10f^2$$

The tessellation frequency required to satisfy the conditions described depends on the size and anisotropy of the molecule. Selection of a suitable value of f is discussed in section 3.1.

2.3. Calculation of Similarity Index. Many forms of similarity index have been previously suggested for the comparison of two sets of values of spatial molecular properties. These include the rank correlation coefficient,[7] the RMS difference,[8] and normalized modifications of the RMS difference such as those suggested by Carbo[9] and Hodgkin.[10]

We have chosen the root mean square difference (RMSd) method as it offers significant computational advantages over other methods. The RMSd similarity is given by

$$\mathrm{RMSd} = [\sum_{i=1,n}\{(P_{A,i} - P_{B,i})^2\}/n]^{1/2}$$

where $P_{A,i}$ and $P_{B,i}$ are the properties of molecule A and B, respectively, at point i, and n is the total number of points.

The RMSd is strictly a measure of *dis*similarity as a minimum in the function represents a maximum in similarity. For a given pair of molecules, this function is calculated for each of the M orientations to be compared but only the best

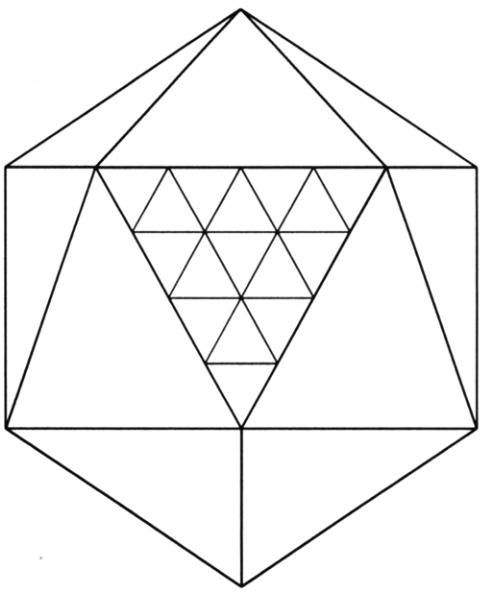

Figure 2. Plan view of an icosahedron showing one face tessellated with a frequency of 4.

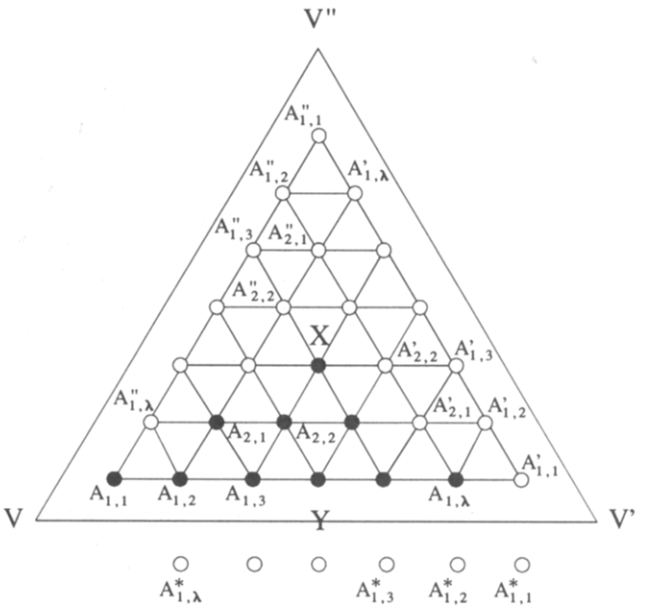

Figure 3. 2D projection of $^1/_{20}$-part of a spherical surface where V, V′, and V″ are adjacent vertices of an icosahedron. All points lie on the surface of the sphere.

m solutions are stored. If at any stage during the summation the value becomes greater than that obtained for the current mth best orientation, the RMSd for the present orientation must be greater than the mth best, and the summation may be prematurely terminated with a consequent saving in computation time.

To obtain maximum advantage from this effect, we introduce two additional features. The first orientation to be compared is that in which the two molecules are aligned such that their principal inertial axes coincide. This gives a high probability of obtaining low RMSd values for the early orientations and, hence, earlier termination of subsequent summations. In addition, the summation is carried out such that the grid points corresponding to the 12 icosahedral vertices are the first to be compared. The vertex, V_i, giving rise to the maximum value of $|P_{A,i} - P_{B,i}|$ is identified, and the remaining points are then included in the summation in order of increasing distance from V_i. This means that the summation is carried

Chart I. Procedure for Generating Approximately Evenly Spaced Points on Surface of Sphere[a]

```
Let V, V' and V" be three adjacent vertices of an icosahedron centre O
Let X be the mid-point of the spherical triangle VV'V"
Let Y be the mid point of the arc VV'
Let χ' be the symmetry operator equivalent to a rotation of +120° about OX[13]
Let χ" be the symmetry operator equivalent to a rotation of -120° about OX
Let χ* be the symmetry operator equivalent to a rotation of 180° about OY
Let A_i,j be the j'th point along the edge of the i'th concentric spherical triangle
Let θ = the nominal grid separation
Let Φ = θ     !initial guess for value of φ
Let r = Φ     !initial range for sampling of φ
DO WHILE Φ has not converged
    DO φ = Φ-r to Φ+r in 20 steps
        Generate point A_1,1 by rotating V along the arc VX by φ
        Generate point A'_1,1 by applying χ' to A_1,1
        Calculate angle ω subtended at O by A_1,1 and X
        Let Λ = nearest integer to (∠A_1,1OA'_1,1 / θ)   !initial guess for value of λ for this value of φ
        DO λ = Λ-1 to Λ+1
            DO face = 1 to INT(1 + λ/3)     ! INT = truncate to integer
                Generate point A_face,1 by rotating A_1,1 along arc VX by (face-1)*3ω/λ
                Generate point A'_face,1 by applying χ' to A_face,1
                Calculate angle α subtended at O by A_face,1 and A'_face,1
                DO step = 2 to λ-3(face-1)
                    Generate point A_face,step by rotating A_face,1 along the arc
                        A_face,1A'_face,1 by (step-1)*α/λ
                ENDDO
            ENDDO
            Generate all points related to A_i,j by operators χ'and χ"
            Generate all points related to A_1,j by operator χ*
            Calculate angles subtended by adjacent points at O
            Let RMSd = RMS difference between these angles and θ
            IF RMSd is lower than the previous best, let λ_opt = λ and φ_opt = φ
        ENDDO
    ENDDO
    Φ = φ_opt
    r = r/10    !Reduce the range over which φ is sampled
ENDDO
Regenerate the set of points using the values of φ_opt and λ_opt
```

[a] The points make up a grid used to effectively scan rotational space.

out over the most dissimilar region of the molecules first, often leading to premature termination at an early stage.

Further computation may be avoided during a search in which the best d hits are required from a database of D molecules. If the value of the summation term exceeds that of the current dth best hit, the summation may likewise be terminated. This leads to savings approaching 100% for the comparison of dissimilar molecules once d 'good' hits have been found and is therefore highly effective when $d \ll D$.

2.4. Sampling Orientation in 3D Space. It is necessary to consider both translational and rotational degrees of freedom when aligning arbitrarily oriented molecules. Optimization of the translational component is approximated by aligning the geometric centers of each of the molecules with the center of the sphere. This approximation is justified in that we are ultimately concerned in studying only those molecules which are very similar to the target molecule.

When performing a database similarity search, it is efficient to initially scan the rotational space of the target molecule and store the corresponding data rather than to scan the rotational space of each database molecule. The gnomonic projection data that must be stored is quite considerable. To

Table I. Characteristics of Polar Grid Designed To Generate Approximately Evenly Spaced Points on Surface of Sphere

nominal grid[a]	grid parameters ϕ[b]	grid parameters λ[c]	points on spherical triangle grid,[d] n_t	mean angular separation[e]	RMS deviation[f]	points on rotation axis,[g] n_r	total points,[h] $n_t \times n_r$
6.0	4.69	10	22	5.92	0.33	60	1320
8.0	6.40	7	12	7.97	0.38	45	540
10.0	8.50	5	7	10.33	0.52	36	252
12.0	9.97	4	5	12.13	0.43	30	150
15.0	12.18	3	4	14.78	0.47	24	96
20.0	15.58	2	2	18.85	1.24	18	36

[a] Requested angular separation of adjacent grid points in degrees. [b] Angle subtended at the center of the sphere by an icosahedral vertex and the nearest point on the grid in degrees. [c] Number of segments into which the edge of the outermost spherical triangle is divided. [d] Number of grid points generated for the two rotational degrees of freedom represented by the surface of the sphere. [e] Mean angular separation of generated adjacent surface grid points in degrees. [f] RMS difference between the angular separation of the generated grid points and the requested separation (or nominal grid) in degrees. [g] Number of points required to sample the third rotational degree of freedom by rotation about the vector from the center of the sphere to the surface point [=360/(nominal grid)]. [h] Total number of grid points required to sample 3D rotational space.

systematically cover the three rotational degrees of freedom of Cartesian space in 10° steps requires 36 × 36 × 36 orientations. However, this introduces a 2-fold degeneracy, and a properly designed scan therefore requires 36 × 36 × 18 orientations.[11] Storing 162 real numbers (f = 4) for each orientation by this brute force method requires approximately 15MB of memory.

Bladon[11] showed how the calculation time and memory requirements could be significantly decreased by the application of symmetry. The icosahedral point group has 60 equivalent positions. If the gnomonic projection (using a set of points with icosahedral symmetry) is calculated for one of these positions, the projections for each of the others may be generated by simply applying the permutation of the vertices corresponding to the appropriate symmetry operator. Thus, to cover all of rotational space, it is only necessary to calculate and store projections for orientations covering $^1/_{60}$ of the same space. Bladon used a Cartesian grid based on the symmetry properties of the icosahedron with ranges of 180°, 72°, and 60° about the x-, y-, and z-axes, respectively. This leads to an unavoidable 2-fold degeneracy and thus, rather than the expected 60-fold improvement, gives a theoretical 30-fold time-saving for the property calculation stage with respect to the brute force method using the same step size. Additionally, there is some degeneracy of points generated at the extremes of the ranges when the symmetry operators are applied. This latter degeneracy can be removed by the use of slightly larger ranges than required by the symmetry properties,[11] but this has the consequence that, after the application of symmetry, the orientations do not uniformly scan 3D rotational space.

An earlier version of the SPERM program used this method proposed by Bladon. We have now improved the efficiency of scanning rotational space by implementing a polar grid to remove the degeneracy. Two degrees of freedom are scanned on a grid covering the spherical surface between three icosahedral points. This could in principle be achieved by triangular tessellation of the icosahedral face. However, angles subtended at the center of the sphere by adjacent points generated by such a method show a significant variation from the mean, and thus the grid does not uniformly scan the two degrees of freedom defined by the surface of the sphere. We use a grid in which the points are equally spaced along the edges of spherical triangles[12] concentric with that defined by three adjacent icosahedral vertices (see Figure 3). This retains the icosahedral relationship between points, and their positions are defined by two parameters ϕ and λ. ϕ is the angle subtended at the center of the sphere (O) by an icosahedral vertex and the nearest point on the grid (i.e., $\angle VOA_{1,1}$ in Figure 3). λ is the number of segments into which the edge of the outermost spherical triangle is divided. These parameters are optimized such that the RMS difference between the angles subtended at the center of the sphere by adjacent points and the required resolution (or nominal grid separation) is a minimum. The effective grid separation is then taken as the mean angle subtended by adjacent grid points. The procedure for generating the points is shown in Chart I.

This method of generating a grid has two major advantages:

It generates approximately evenly spaced points on the surface of a sphere.

The application of icosahedral symmetry operators results in no degeneracy (except when the last point lies at the center of the spherical triangle where it has 3-fold degeneracy).

The grid described covers only two rotational degrees of freedom, scanning the space in which an arbitrary axis of the molecule passes through all points on the surface of the sphere. The third degree of freedom is scanned by rotation about this axis. Only one-third of the points within an icosahedral face need be sampled (those marked ●) as the others (in this face and the 19 other faces) are generated by the 60 symmetry operators. An additional advantage of the use of icosahedral symmetry is that for each symmetry operator there is a corresponding improper one, the application of which results in the generation of the structural enantiomer. Thus, by using an additional 60 vertex permutations we are also able to sample the rotational space of the enantiomer. The properties of such a grid are summarized in Table I.

Use of such a grid offers considerable savings in the computation time and memory requirements for the calculation of the molecular properties. The coverage of 3D rotational space with a nominal 10° grid spacing requires the calculation of a gnomonic property projection for each of 23 328 (36 × 36 × 18) orientations by a brute force method, 1197 (19 × 9 × 7) orientations by Bladon's method, and only 252 (7 × 36) by this method (see Table I).

2.5. Optimizing Orientation in 3D Space. The objective of a database similarity search is to rank the database molecules according to their similarity to the target structure. We therefore wish to find only the single orientation of each database molecule which shows the highest similarity to the target molecule (i.e., the global minimum in the RMSd function). In a more detailed study of a single pair of molecules

Table II. Functions of Principal Moments of Inertia of Selected Molecules

molecule	M1[a]	M12[b]	M23[c]
n-hexane	129	1.039	5.30
n-nonane	372	1.022	10.12
n-dodecane	818	1.013	16.91
adamantane	90	1.000	1.00
dodecahedrane	200	1.000	1.00
benzene	48	2.000	1.00
chair cyclooctane	117	1.528	1.05

[a] First principal moment of inertia [all atoms with unit mass]. [b] Ratio of first and second principal moments of inertia. [c] Ratio of second and third principal moments of inertia.

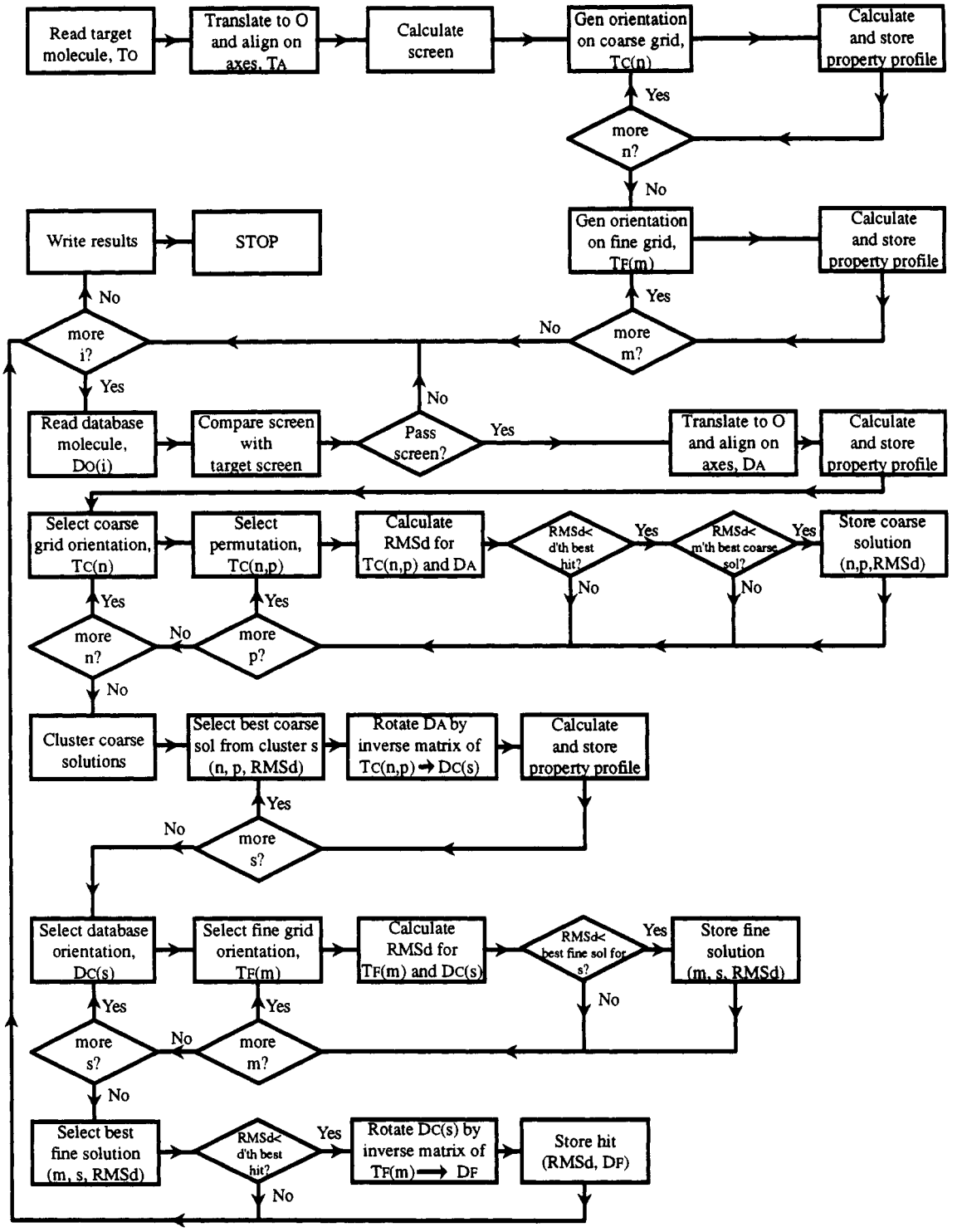

Figure 4. Flow chart showing the principal steps of the SPERM method.

we may be interested in finding a series of minima. These different objectives give rise to differing approaches to the location of the required minima. Dean and Chau[14] carry out partial minimizations from a number of random starting orientations. The resulting orientations are then clustered, and one member of each cluster is subjected to full optimization. Such a method is very demanding of processing time and is unsuitable for database searching. In our scheme, we initially carry out a systematic scan of rotational space as described in section 2.4 using a coarse grid (generally 20°). The orientation having the minimum RMSd value at this grid separation may not lie in the same well in orientational space as the global minimum. We therefore save the best m solutions (usually 10) and cluster them on the basis of the similarity of the rotation matrices corresponding to the coarse grid target molecule orientations. By selecting the member of each cluster with the minimum RMSd value, we obtain a number of solutions lying in discrete wells in the RMSd function.

The inverse of the rotation matrix corresponding to target molecule orientation of each of these solutions is then applied to the coordinates of the database molecule. This generates a set of orientations of the database molecule corresponding to local minima in the RMSd function with respect to the reference orientation of the target molecule. The gnomonic projection is recalculated for each member of this set. Each is then compared with a number of orientations of the target molecule covering only the rotational space in the region of the reference orientation. A fine grid for two degrees of

rotational freedom is generated in a similar manner to that for the coarse grid. However, it only covers that part of the spherical surface within the 'Arctic circle' where the 'latitude' is greater than $90-\theta/2$ deg (θ is the nominal angular separation of the coarse grid). The nominal grid separation (θ') of this fine grid is generally set to about $\theta/5$. The third degree of freedom is then scanned in the same way as for the coarse grid by rotation about the axis from the center of the sphere through the point on the surface grid but over a range of $-\theta/2$ to $+\theta/2$ in steps of θ'. The comparison is made in an efficient manner as the gnomonic projections for these fine grid orientations of the target molecule are precalculated and stored at the beginning of the run. The database molecule orientation giving rise to the lowest RMSd (the global minimum) is finally rotated by the inverse of the matrix associated with the corresponding fine grid target molecule orientation and the solution stored.

2.6. Database Implementation. It was mentioned earlier that the performance of database searching could be improved by requesting that only the best d hits be retained, thus allowing premature termination of the summation term (see section 2.3) if its value exceeded that of the current dth best hit. Performance may be further improved by implementing a screening system in which database molecules showing a low similarity to the target molecule may be rejected without applying the SPERM algorithm at all. However, the type of screening method required is fundamentally different in concept to one which would be used where a database is searched for entries which exactly match a given query (even though the query may be defined with certain tolerances), as for 2D substructure searching[15] or 3D pharmacophore searching.[16] In the latter case, the properties being searched are functions of one molecular structure only. Examples might be the presence of a O···N nonbonded distance in the range 5–7 Å or the presence of two hydrogen bond acceptors in the molecule. In such cases, it is possible to set screens which show a direct correspondence to the properties being searched, and thus entries may be eliminated if they cannot possibly match the query. In the case of our similarity search, we are not looking for exact matches, and the property being searched is a function of both the target and the database molecule. It is therefore not possible to store screen data which show a direct correspondence to the similarity of a database molecule to a target molecule which is unknown at the time of screen calculation. Hence, screening has to be based on properties calculated for individual molecules which, if matched between the target and database molecule, are indicative of high similarity between them. We currently use four properties in the database screen.

The molecular volume (VOL)

The first principal moment of inertia [with unit atomic masses] (M1)

The ratio of the first and second principal moments of inertia (M12)

The ratio of the second and third principal moments of inertia (M23)

The significance of the latter three properties may be understood by considering the values in Table II. If a molecule is oriented such that its first principal inertial axis is aligned with the z-axis, the second with the y-axis, and the third with the x-axis, the value of M1 is related to the length of the molecule (along the x-axis), M12 to the degree to which a 3D shape has been squashed ($1 \leq M12 \leq 2$) toward the xy-plane, and M23 to the degree to which it has been squashed ($M23 \geq 1$) toward the xz-plane.

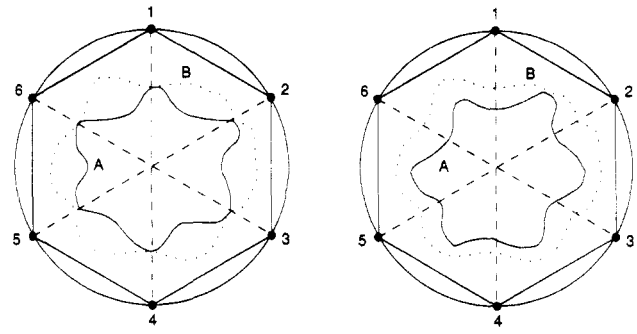

Figure 5. 2D analogy of gnomonic projection showing how the projected property values for molecule A depend on its orientation with respect to the projection points. Molecule B is shown in the orientation in which the RMSd similarity function with respect to A has been optimized in each case.

The values of each property are calculated for a representative cross-section of the database. For each property, the list of values is sorted and divided into 16 equally populated 'bins'. The values corresponding to the boundaries between adjacent bins are then stored. The properties are then calculated for all molecules in the database, each is assigned to the appropriate bin, and the corresponding bit is set in a 16-bit word. Thus, each molecule is represented by four 16-bit words, each of which has a single bit set.

At the beginning of a database search, the same property calculation is carried out for the target molecule and the appropriate bit set in each of the target property screens. A tolerance (or window) is allowed by additionally setting b bits to either side. For each database molecule, each of the four screens (D_i) is then compared in a bitwise manner with the corresponding target molecule screen (T_i), and the structure is rejected unless the following equality holds for *all* of the screens:

$$\forall_{1 \leq i \leq 4} \; (T_i).\mathrm{AND}.(D_i) = D_i$$

Because of the imprecise relationship between the matching of the screen properties and the measured RMSd similarity, there is a small but finite chance that a molecule which shows a high similarity to a given target may be rejected by this screening method. It is therefore important that appropriate screen properties are selected and a sufficiently large window is applied to keep this identification of false negatives to a minimum. Some results of the application of database screening are presented and discussed later.

2.7. Summary. The principal steps in the implementation are summarized in the flow chart shown in Figure 4.

3. RESULTS AND DISCUSSION

3.1. Selection of Optimum Number of Projection Points. Increasing the number of projection points improves the resolution of the property profile but leads to an increase in computation time. Hence a balance must be found between validity and speed. We have found that the use of a small number of sampling points can lead to misleading results. This can be appreciated by considering the 2D analogy depicted in Figure 5. Figure 5 (left) shows an orientation of molecule B which has been optimized with respect to a fixed reference orientation of A, giving a solution with an RMSd of 0.0. If A is given a different fixed orientation as in Figure 5 (right), a different solution is obtained with a larger RMSd value, although it is clear that this would be a better solution if more

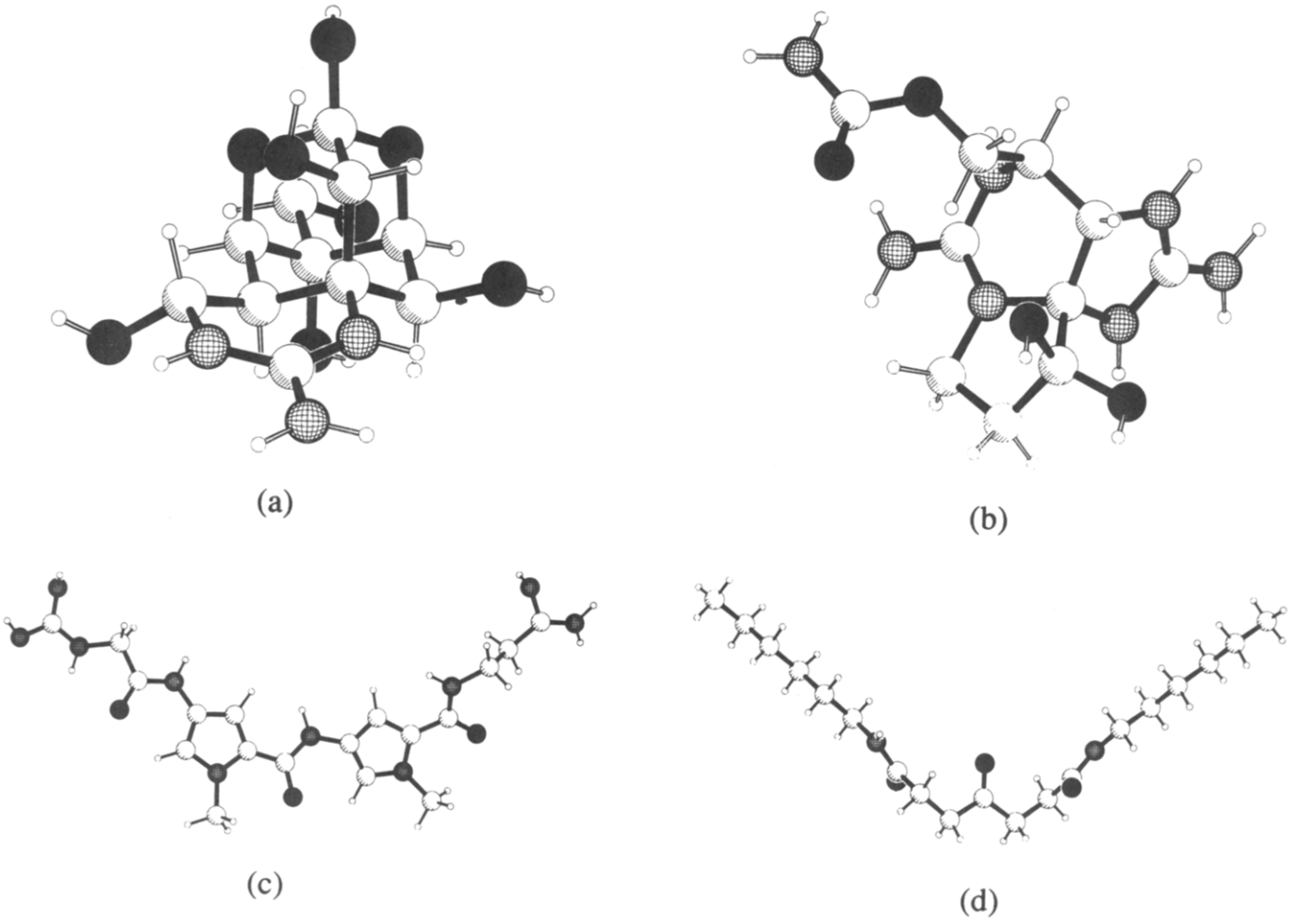

Figure 6. Crystal structures of molecules used in the study: (a) tetrodotoxin; (b) saxitoxin; (c) netropsin; (d) *N,N'*-di-*n*-hexyl-4-oxoheptanediamide.

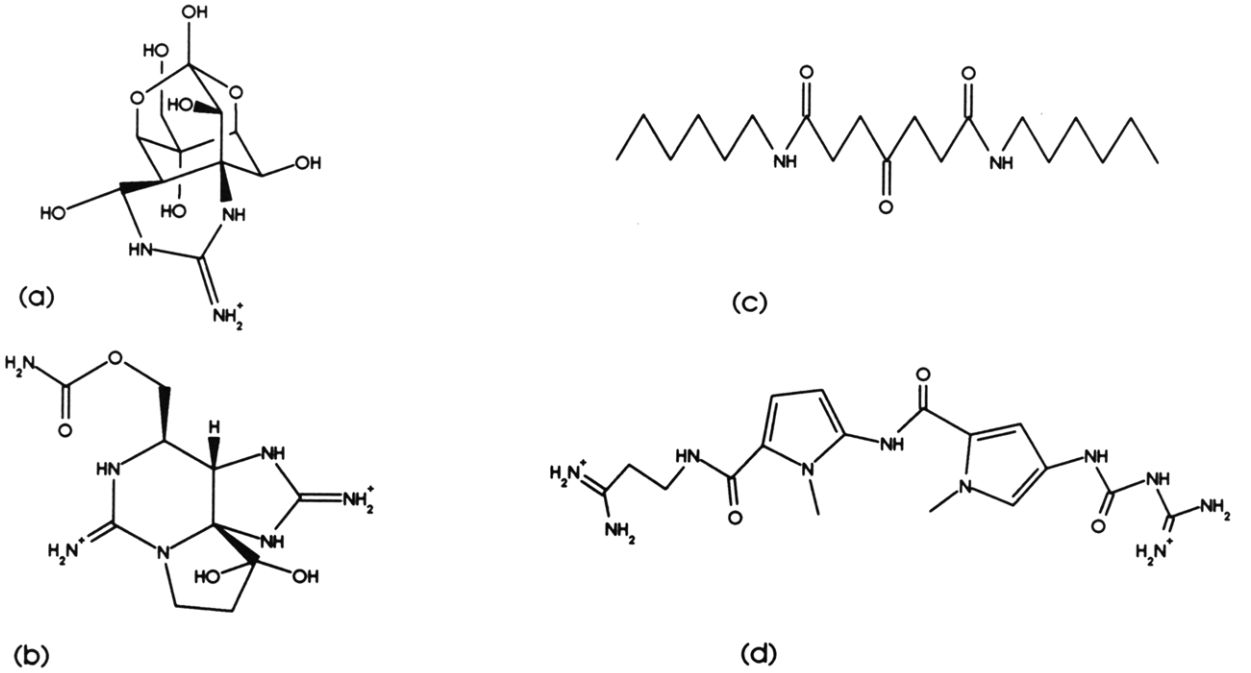

Figure 7. Chemical structures of molecules used in the study: (a) tetrodotoxin; (b) saxitoxin; (c) netropsin; (d) *N,N'*-di-*n*-hexyl-4-oxoheptanediamide.

surface points had been considered. This is a consequence of the inadequate sampling of the surface profile with a small number of points.

The dependence of this effect on the number of projection points was studied in the following way. The orientation of a molecule A was optimized with respect to a number of random orientations of a molecule B. With an infinite number of projection points, the RMSd values of the optimized solution in each case should be the same. Thus, the standard deviation of these values can be taken as a quantitative measure of the

Table III. Dependence of Optimized Similarity Index for Pair of Molecules on Reference Orientation of Fixed Molecule

molecule pair	tessellation frequency, f	points	RMSd similarity index (for 100 orientations) min	mean	max	SD
TTX/STX[a]	6	362	0.570	0.573	0.578	0.0021
	5	252	0.568	0.572	0.579	0.0031
	4	162	0.568	0.574	0.583	0.0031
	3	92	0.557	0.577	0.595	0.0076
	2	42	0.518	0.550	0.615	0.0198
	1	12	0.253	0.514	0.863	0.1279
NET/HEP[b]	6	362	0.546	0.552	0.568	0.0035
	5	252	0.544	0.552	0.571	0.0041
	4	162	0.545	0.553	0.568	0.0038
	3	92	0.544	0.553	0.569	0.0047
	2	42	0.536	0.559	0.576	0.0093
	1	12	0.352	0.435	0.585	0.0414

[a] TTX, tetrodoxin; STX, saxitoxin. Examples of molecules showing isotropic shape. [b] NET, netropsin; HEP, *N,N'*-di-*n*-hexyl-4-oxoheptanediamide. Examples of molecules showing anisotropic shape.

effect. The molecules tetrodotoxin (TTX) and saxitoxin (STX) were used as one pair of test molecules as these have been extensively studied in other studies.[7,8,11,14,17] As can be seen in Figure 6, they are rather isotropic molecules, and thus a second pair of molecules showing greater anisotropy was also selected. Netropsin (NET), which has been used by us in a previous study,[5] was compared with *N,N'*-di-*n*-hexyl-4-oxoheptanediamide (HEP), a molecule identified by a database search as showing high similarity to NET. The chemical structures of all four molecules are shown in Figure 7. In each case, the molecules were fitted by the SURDIS property using a coarse grid separation of 10° followed by a fine grid separation of 2°. The fitting was then repeated for 100 random orientations of the second molecule and the procedure carried out for tessellation frequencies between 1 and 6. The results are summarized in Table III.

In both cases it can be seen that when f lies between 4 and 6 the standard deviation of the RMSd scores is rather small and constant. This in fact corresponds to the variation due to the 2° grid separation of the search. For $f \leq 3$, the standard deviation increases and the likelihood of 'false' optima increases. In the context of a database search these figures are very significant. For a reported score x, the 95% confidence limits are $x \pm 1.96\sigma$. This has the consequence that, for a search in which TTX is the target molecule and $f = 2$, we can only be confident that the top 25 hits genuinely lie in the top 100. The situation is somewhat better for NET where the corresponding figure is 83. For $f \geq 4$, the corresponding figures are 85 for TTX and 90 for NET. We conclude, therefore, that great care should be taken in the interpretation of results for $f < 4$ (i.e., fewer than 162 projection points). In particular, the 32 points (from the vertices of an icosahedron and a dodecahedron) used by Bladon and in earlier versions of SPERM would appear to be inadequate for the production of reliable and reproducible similarity indices. There appears to be no advantage in using $f > 4$ (at least for a 2° grid separation), leading us to conclude that $f = 4$ (162 points) offers the best compromise between validity and computation time.

A more even distribution of projection points over the surface of the sphere could be generated in the manner described in section 2.4 for sampling the rotational space. However, for a given number of projection points, it is not important that they cover the surface of the sphere as evenly as possible, merely that they sample as much of the molecular surface as possible and retain an icosahedral symmetry relationship. Therefore, we currently employ the method of simple tessellation.

Table IV. Percentage Saving in Time by Premature Termination of Summation Term in RMSd Similarity Function for Number of Pairs of Molecules

	NET[a]	HEP[b]	TTX[c]	STX[d]	BEN[e]	PYR[f]
NET	79.0					
HEP	78.3	80.4				
TTX	16.5	19.5	69.9			
STX	21.1	25.1	60.9	73.9		
BEN	4.8	4.9	5.0	7.0	85.7	
PYR	4.5	4.7	4.9	6.9	83.9	82.4

[a] Netropsin. [b] *N,N'*-Di-*n*-hexyl-4-oxoheptanediamide. [c] Tetrodotoxin. [d] Saxitoxin. [e] Benzene. [f] Pyridine.

3.2. Computation of RMSd Similarity Function. It was stated in section 2.3 that the comparison of a fixed orientation of one molecule with many orientations of another could be made more efficient if we are only interested in the best few solutions. Table IV shows the percentage time saved with respect to carrying out the full summation for all orientations for the comparison of a variety of molecules. The rotational space was scanned on a coarse grid of 20°, giving a total of 4320 orientations (36 points × 60 permutations × 2 enantiomers), and the best 10 solutions were retained in each case.

For the comparison of similar molecules, the saving is in the range 70–90%, falling to only about 5% for dissimilar molecules. This can be understood in terms of the greater difference between maximum and minimum calculated similarities for different relative orientations of similar molecules than for dissimilar ones. Hence, database searches will be significantly accelerated by implementation of methods which speed the processing of dissimilar molecules as described in section 2.6.

3.3. Database Searches. All of the searches reported here have been carried out on a database derived from the Cambridge Structural Database (CSD), selecting only organic compounds which show no disorder. By matching the stored tables of chemical connectivity and crystal connectivity, we have assigned bond orders and added missing hydrogens (using standard bond lengths and angles) to the crystal structure. This gives us a database of 33 123 compounds showing a wide diversity of chemical structure.

Searches were carried out on this database using TTX and NET as target molecules. The SURDIS property was used with a 162-point projection ($f = 4$). A comparison was made over a coarse grid of 20° for each molecule (including enantiomers), and the 10 best orientations were saved and clustered. The best member of each cluster was then compared with the target molecule on a fine grid of 4°. This procedure was repeated for each database molecule, and the top 100 solutions were saved and reported. In both cases the search time on a VaxStation 3100 M76 running under VMS 5.4 (~7 VUPs) was between 6 and 7 h.

The TTX search produced hits with RMSd values in the range 0.395–0.475 Å, compared with a range of 0.556–1.002 Å for the NET search. Figure 8a shows three orthogonal views of the top 10 hits from the TTX database search superimposed on the van der Waals surface of TTX. Figure 8b shows the equivalent superposition for NET. It can be clearly seen that the method is effective in identifying database molecules showing close shape similarity to the target molecule. That the hits exhibit a wide diversity of chemical structure can be seen by inspection of the chemical structures of the top

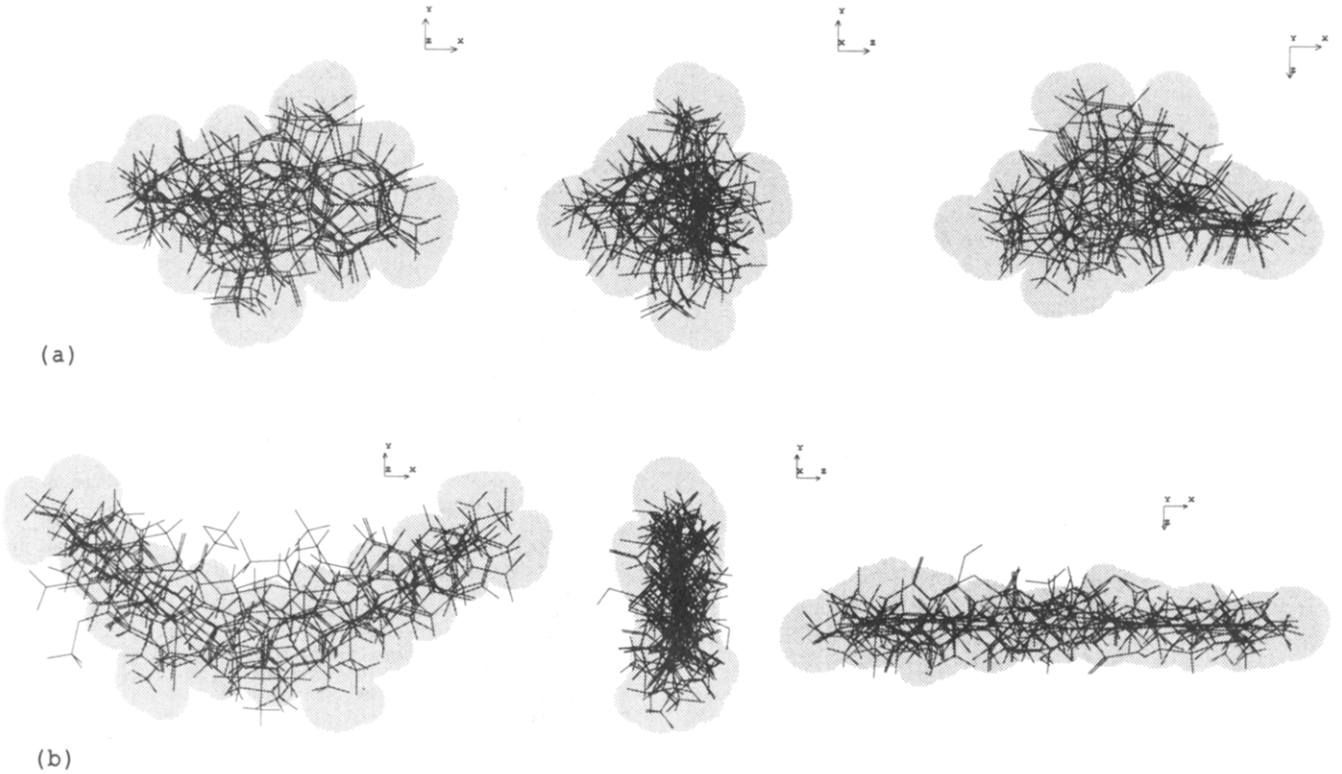

Figure 8. Orthogonal views of the top 10 hits from two database similarity searches superimposed on the van der Waals surface of the target molecule. (a) Tetrodotoxin search. (b) Netropsin search.

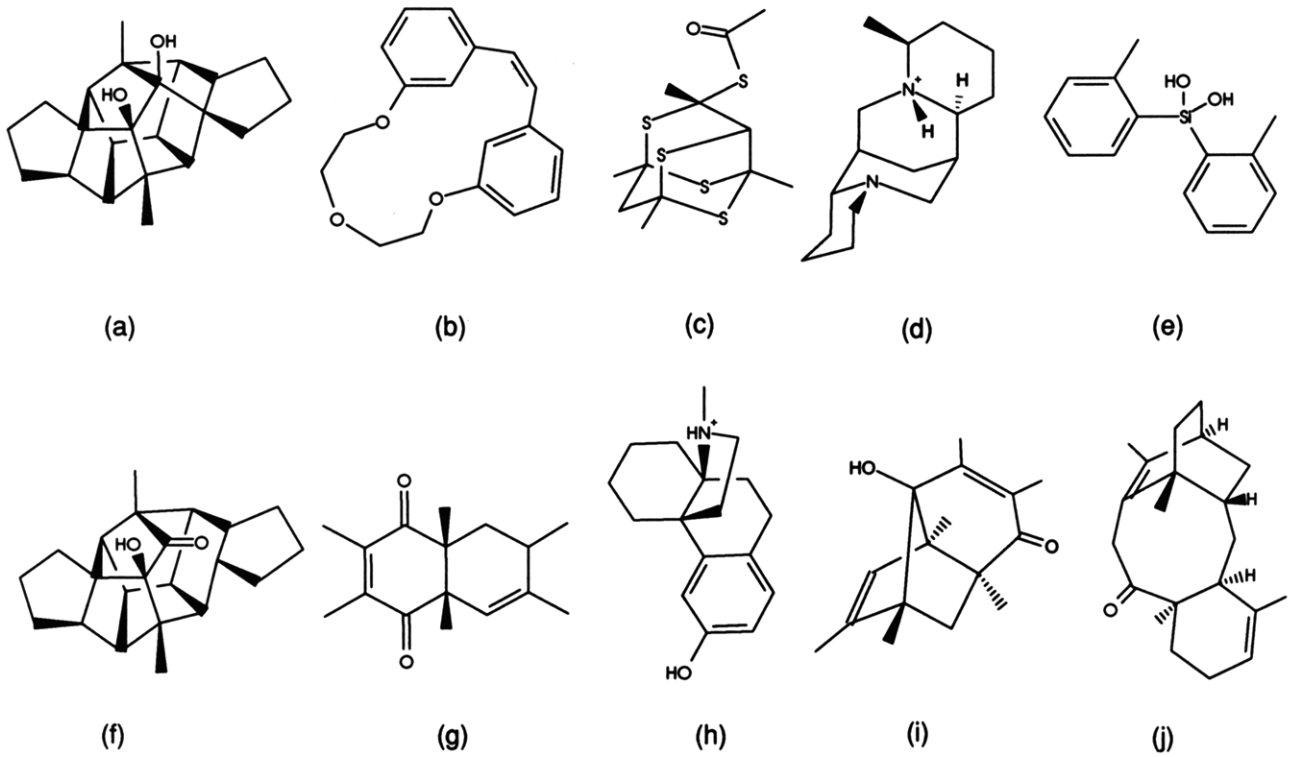

Figure 9. Chemical structures of the top 10 hits from the tetrodotoxin similarity search. CSD REFCODEs of the structures are as follows: (a) DODECB; (b) CEYJUL; (c) DOZKOS; (d) MSPRTC; (e) VAFWII; (f) DODECC; (g) HMTHNQ10; (h) HMHASB; (i) HYMDDO; (j) BUYLEM10.

10 hits from the TTX search in Figure 9. We showed in our previous paper[5] that in addition to molecules showing high chemical similarity to the target, the method also identified molecules which were chemically different but had been reported as showing similar biological activity.

3.4. Database Screening. The consequences of database screening may be modified by adjusting the bit screen tolerance (or window) applied. A small window leads to a high rejection rate (screenout) but leads to the rejection of some of the compounds which would have produced a low RMSd score (i.e., high similarity). A larger window reduces this risk but is less effective at rejecting compounds of low similarity. The results of database searches using different window sizes are summarized in Table V.

Table V. Effect of Screening on Database Searches

target molecule	window[a]	screen-out, %	time, min	time saving, %	no. lost from top 100
NET[b]	none	0.0	369		0
	6	78.5	146	60.4	6
	4	93.1	98	73.4	31
TTX[c]	none	0.0	413		0
	6	78.8	167	59.6	9
	4	91.7	122	70.5	20

[a] Bit screen tolerance in bits. [b] Netropsin. [c] Tetrodotoxin.

A window of 4 leads to the loss of a large number of structures from the top 100. Most of these are rejected on the M12 screen, which appears to have a lower predictive value for similarity than the other screens. A window of 6 leads to the rejection of fewer hits (all of which are rejected on M12) and offers a better compromise. We are currently considering alternative screen properties which might give higher screenout with a lower rejection of hits.

It can be seen that the percentage saving in time is smaller than the percentage screenout. This is due to the fact that when no screening is applied, the database molecules of low similarity are processed rather faster than those of high similarity due to earlier termination of the summation stage. However, significant savings in time may be made by the application of screening, and this is particularly important when large databases are being searched. The use of a small window also allows for a database to be quickly searched in order to obtain a 'feel' for the results before a more exhaustive search is made.

4. CONCLUSIONS

We have developed a method to rank molecules in a 3D structural database according to their similarity in a 3D property profile with respect to a target molecule. The SPERM program uses gnomonic projection of molecular properties and application of icosahedral symmetry to a polar grid in the comparison process to search all of rotational space in a highly efficient manner. With the molecular property projected onto the vertices of a tessellated icosahedron, at least 162 points are required to produce reliable and reproducible similarity indices. Database search times may be significantly improved by the application of screening, but the advantage is smaller than that typically found by applying screens in traditional database searching for exact matching of properties. The method works equally well for molecules having isotropic and anisotropic properties, but the current implementation does not search either translational or conformational space.

The SPERM program may be used to search databases for potential lead compounds in drug design. While a quantitative measure of similarity to the target structure is obtained for each database entry, it cannot be taken as a measure of its potential as a lead compound. Different similarity measures produce different rankings, and there is no independent way to judge which is superior. However, searches have been able to identify molecules of diverse chemical structure which show similar biological activity to the target structure. This gives us confidence that other chemically diverse molecules identified by this method may be interesting lead compounds for further study.

REFERENCES AND NOTES

(1) Johnson, M. Similarity-based methods for predicting chemical and biological properties: a brief overview from a statistical perspective. In *Chemical Information Systems*; Bawden, D., Mitchell, E. M., Eds.; Ellis Horwood: Chichester, England, 1990; pp 149–159.

(2) Martin, Y.; Bures, M. G.; Willett, P. Searching Databases of Three-Dimensional Structures. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers: New York, 1990; pp 213–263.

(3) See, for example: (a) Pepperrell, C. A.; Willett, P. Techniques for the calculation of three-dimensional structural similarity using inter-atomic distances. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 455–474. (b) Manaut, F.; Sanz, F.; José, J.; Milesi, M. Automatic search for maximum similarity between molecular electrostatic potential distributions. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 371–380.

(4) For example, see: (a) Doweyko, A. M. The Hypothetical Active Site Lattice. An Approach to Modeling Active Sites from Data on Inhibitor Molecules. *J. Med. Chem.* **1988**, *31*, 1396–1406. (b) Kato, Y.; Itai, A.; Iitaka, Y. A Novel Method for Superimposing Molecules and Receptor Mapping. *Tetrahedron* **1987**, *43*, 5229–5236.

(5) van Geerestein, V. J.; Perry, N. C.; Grootenhuis, P. G.; Haasnoot, C. A. G. 3D Database Searching on the Basis of Ligand Shape Using the SPERM Prototype Method. *Tetrahedron Comput. Methodol.* **1990**, *3*, 595–613.

(6) Meyer, A. Y.; Richards, W. G. Similarity of Molecular Shape. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 427–439.

(7) Chau, P. L.; Dean, P. M. Molecular recognition: 3D surface structure comparison by gnomonic projection. *J. Mol. Graphics* **1987**, *5*, 97–100 (and pp 88–89 for color plates).

(8) Dean, P. M.; Callow, P.; Chau, P.-L. Molecular recognition: blind searching for regions of strong structural match on the surfaces of two dissimilar molecules. *J. Mol. Graphics* **1988**, *6*, 28–34.

(9) Carbo, R.; Leyda, L.; Arnau, M. How Similar Is a Molecule to Another? An Electron Density Measure of Similarity Between Two Molecular Structures. *Int. J. Quantum Chem.* **1980**, *17*, 1185–1189.

(10) Hodgkin, E. E.; Richards, W. G. Molecular Similarity Based on Electrostatic Potential and Electric Fields. *Int. J. Quantum Chem., Quantum Biol. Symp.* **1987**, *14*, 105–110.

(11) Bladon, P. A rapid method for comparing and matching the spherical parameter surfaces of molecules and other irregular objects. *J. Mol. Graphics* **1989**, *7*, 130–137.

(12) A spherical triangle is that portion of the surface of a sphere bounded by the arcs joining three points on the surface.

(13) A clockwise rotation about a vector **AB** when looking from A to B is taken to be positive.

(14) Dean, P. M.; Chau, P.-L. Molecular recognition: optimized searching through rotational 3-space for pattern matches on molecular surfaces. *J. Mol. Graphics* **1987**, *5*, 152–158.

(15) For example, see: (a) Feldman, A.; Hodes, L. An Efficient Design for Chemical Structure Searching. I. The Screens. *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 147–152. (b) Dittmar, P. G.; Farmer, N. A.; Fisanick, W.; Haines, R. C.; Mockus, J. The CAS ONLINE Search System. 1. General System Design and Selection, Generation, and Use of Search Screens. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 93–102.

(16) For example, see: (a) Jakes, S. E.; Willett, P. Pharmacophoric pattern matching in files of 3-D chemical structures: selection of inter-atomic distance screens. *J. Mol. Graphics* **1986**, *4*, 12–20. (b) Murrall, N. W.; Davies, E. K. Conformational Freedom in 3-D Databases. 1. Techniques. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 312–316.

(17) Namasivayam, S.; Dean, P. M. Statistical method for surface pattern-matching between dissimilar molecules: electrostatic potentials and accessible surfaces. *J. Mol. Graphics* **1986**, *4*, 46–50.