Quadratic Shape Descriptors. 1. Rapid Superposition of Dissimilar Molecules Using Geometrically Invariant Surface Descriptors

Brian B. Goldman and W. Todd Wipke*

Molecular Engineering Laboratory, Department of Chemistry and Biochemistry, University of California, Santa Cruz, California 95064

Received July 17, 1998

In this paper, we present a novel approach to shape-based molecular similarity searching. The method that we introduce is able to superimpose dissimilar molecules by using geometrically invariant molecular surface descriptors. The shape descriptors are calculated by least-squares fitting of a quadratic function to small sections of the molecular surface of a ligand. Invariant geometric properties of the approximated surface patch are then extracted from the fitted quadratic function. The extracted properties are used to quantify the shape and to obtain a canonical orientation for this section of surface. The superimposition algorithm uses these geometric invariants to recognize similar regions of surface shape existing on two molecules and to bring these regions (and consequently the molecules) into registration. Because these geometric descriptors are based upon local surface shape, the superimposing algorithm is insensitive to the connectivity and the relative sizes of the molecules being matched. The capabilities of our algorithm are demonstrated by superimposing dissimilar ligands known to inhibit the same enzyme system. In all cases examined the algorithm generates superpositions that are in agreement with crystallographic results. The algorithm is also applied to align the two different proteins on the basis of the shape of their active sites.

INTRODUCTION

The ability to discover similarity between different molecules is an important aspect in the drug design process.[1,2] Similarity between molecules occurs when the molecules have some attribute in common. Recognizing this similarity is often a first step in scientific theory formulation. For example, the formulation of a pharmacophore hypothesis involves recognizing the structural similarities among a series of compounds sharing the same biological activity.[3−6] Other applications of molecular similarity include three-dimensional database searching,[7−12] and alignment of active compounds in preparation for comparative molecular field analysis.[13]

Structural features used in assessing molecular similarity may be derived from the graph-theoretical aspects of the structure, e.g., the presence of functional groups, rings, atom partial charges, MACCS structural keys,[14,15] and a variety of connectivity indices.[16−20] The availability of good modeling algorithms and experimentally determined three-dimensional structures by crystallography and nuclear magnetic resonance (NMR) now facilitates using structural features derived from the geometry of molecules to determine similarity.

From the three-dimensional structure, features such as distances or angles between atoms, bond, or rings may be derived and tabulated as a distribution. Structures with similar distributions may be said to be similar.[21−24] Since distributions can be compared quickly, these features make good screens for similarity searching over a large database of three-dimensional structures.[25] Although this method finds molecules with similar distributions, it does not provide an atom−atom correspondence or three-dimensional spatial alignment of molecules. As indicated by Wild and Willett, these features are not based on the molecular properties (e.g., electrostatic, hydrophobic, or steric fields) that are associated with contemporary methods for predicting biological activity.[10] Thus, these methods derive a three-dimensional structural similarity that is independent of the orientation of the molecules being compared.

The most interesting similarity measures are based on superimposing one molecule on another such that key structural features on one molecule coincide with or are close to identical or similar features on the other. These features can be heteroatoms, aromatic rings, or hydrogen bond donor/acceptor atoms typical of pharmacophore models. Alternatively, the features can be surface shape descriptors, in which case the alignment would be based on steric surface shape. In either case, the molecular alignments are interesting because they suggest a correspondence between two molecules that might imply function, i.e., how one molecule might mimic another. In addition, the correspondence enables a chemist to consider further structural modifications that might enhance the structural similarity. Finally, the alignment in three dimensions permits evaluation of the types of molecular properties[10] (e.g., electrostatic, hydrophobic, or steric fields) that are associated with contemporary methods for predicting biological activity. Similarity that is based on three-dimensional alignments provides much more chemically useful information than is provided by other types of similarity measures.

Determining a superimposition of two rigid molecules in order to maximize a similarity metric that is dependent upon the relative orientations of the molecules is a combinatorial problem in a six-dimensional space (three degrees each for

* To whom correspondence should be addressed. E-mail: wipke@chemistry.ucsc.edu.

10.1021/ci980213w CCC: $19.00
Published on Web 05/02/2000

rotation and translation). A systematic search to maximize the surface overlap between two rigid drug sized molecules can require days of CPU time on a serial computer.[26] The vast size of this search space has prompted many researchers to approach the molecular similarity problem with stochastic optimization algorithms (simulated annealing, genetic algorithms), that are well-suited to searching large problem spaces.[6,10,27,28] While some implementations of these stochastic algorithms are robust in that they allow the molecules being compared conformational freedom during the matching procedure,[6] they are generally too demanding of computational resources to be used for similarity searching over databases of nontrivial size. A notable exception is the genetic algorithm developed by Thorner for the comparison of molecular electrostatic potentials.[29] Furthermore, the nondeterministic nature of these algorithms exacerbates their computational expense by necessitating repeated runs on identical inputs in an attempt to avoid convergence to suboptimal solutions. These stochastic algorithms are useful when computational speed is not a primary concern, for example, in the generation of pharmacophore hypothesis from a series of active compounds.[6]

An alternative to using stochastic search methods is to develop deterministic algorithms capable of reducing the search space to a computationally manageable size. When using deterministic algorithms to ascertain steric molecular similarity, the crux of the problem becomes how to represent molecular shape so that important shape features can be readily identified. Of all the three-dimensional features, molecular shape seems the most difficult to characterize. Heteroatom location is easy to characterize by a Cartesian coordinate system and atom type. Hydrogen bond donor/acceptor groups have an additional vector component. However, characterizing molecular shape has remained a difficult problem.

Several methods for shape representation that afford efficient molecular similarity searching include gnomonic projection of molecular surfaces,[30−33] molecular principal moment analysis,[8] and Fourier descriptors.[34] Each of the above techniques uses an abstract representation of molecular shape to afford efficient similarity searching. However, they also encode global shape information. This fact makes the algorithms using these representations sensitive to the global characteristics (e.g., size) of the molecules being compared. For example, it has been demonstrated that the gnonomic projection method, which assesses molecular similarity by the projection of molecular properties onto the verticices of a polyhedron surrounding two molecules, is sensitive to the location of the polyhedron center.[35] Techniques that use global characteristics of molecules in their shape representations are biased toward finding molecules with globally similar properties. Therefore, while these algorithms are well-suited for finding pairs of molecules that are similar in both size and shape, they are inappropriate for *subshape similarity searching.*

Subshape similarity searching locates the common shape features shared by two or more molecules. For example, if two molecules are of greatly different size yet contain a region of surface that is of common shape, then subshape similarity searching can be used to recognize this shape feature. The two molecules may then be aligned such that the common surface regions are superimposed. Potential applications of subshape similarity searching include searching for small molecule mimics of a protein.

One technique for subshape similarity searching is the molecular skin matching algorithm of Masek et al.[36] Using molecular surfaces with thickness (molecular skins), coupled with a minimization algorithm, Masek et al. were able to determine orientations of molecules that maximize their surface overlap. Because the algorithm is maximizing the surface area in common between two molecules, it is able to find superpositions that emphasize common shape features. The algorithm is therefore applicable to subshape similarity matching. Using this algorithm, Masek et al. were able to correctly align two inhibitors of the protein elastase that are substantially different in size. One ligand is a 56 residue peptide and the other a small difluoroketone. However, due to the analytical nature of the minimization algorithm, the run time was exceptionally high (12 CPU days on an SGI-R4000). While Perkins et al.[28] used the same definition of molecular skins, they were able to substantially reduce the run time of the procedure by using a numerical representation of the molecular skin and a simulated annealing algorithm as the optimization protocol. Their method also reproduced experimentally verified molecular superimpositions. Both Masek et al. and Perkins et al. showed that when the surfaces are aligned, the molecules are aligned, which indicates that the surface contains important information; however, they only used this information as an evaluation function *after* they had generated a trial alignment. The goal of our research was to use surface shape to determine the alignment directly.

Our work has been aimed at developing a simple efficient shape descriptor. This paper describes our shape descriptor and its use in the shape similarity of rigid three-dimensional structures. Applications of this work to shape complementarity[37,38] and docking of rigid[39] and of flexible structures[40] is reported elsewhere.

CATEGORIES OF SHAPE DESCRIPTORS

Surface versus Bulk. A molecular surface description pertains to the van der Waals outer surface, an isopotential surface, a Connolly solvent accessible surface, or an electron isodensity surface. The molecular surface may further be smoothed by a variety of functions, including wavelets,[41] Fourier descriptors,[42] spherical harmonics,[43] or B-splines.[44] We differentiate between a description of the surface shape and a description of the bulk volume contained within the surface.

A bulk shape description pertains to characterizing a molecular shape by what objects or object−distance−object tuples can fit inside. The contained objects may or may not be tangent at one or more places to the molecular surface. Aligning two molecules by bulk description may assume the largest common volume, but not the largest common surface similarity. We felt that molecular surface was the more important aspect of shape and focused on that.

Local versus Global. A local shape description can only reflect a local region of the surface or bulk as the case may be. A global descriptor typically measures the maximum dimensions of the molecular shape, e.g., length, width, height. Local descriptions are useful in subshape and full shape similarity, whereas global descriptors are only useful in full

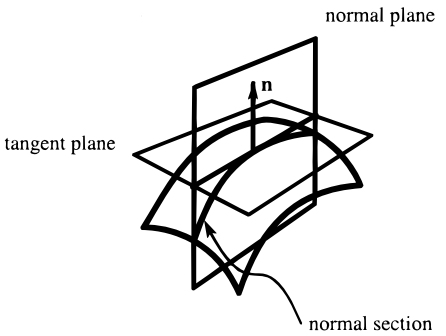

Figure 1. The tangent plane, normal plane, and normal section of a surface patch.

(global) shape similarity. For this reason, we focused on developing a local shape descriptor.

METHODOLOGY

We now describe, in detail, the geometric basis of our shape descriptor and how it is used for surface shape matching. Following that, we present the algorithm for performing surface matching along with an analysis of its complexity and the experimental details.

The cornerstone of our algorithm is the geometrically invariant representation of small sections of molecular surface. Our method is predicated upon the observation that if two molecules have similar surface shape over a large region of their molecular surfaces, then smaller sections of their surfaces must also be similar. Therefore, the alignment of local surface patches provides a mechanism for uncovering larger regions of shape similarity. If small sections of molecular surface can be described in a geometrically invariant, compact, and precise manner, then efficient shape matching can be realized. This can be accomplished by aligning small sections of surfaces from two different molecules and subsequently checking for larger regions of shape correspondence.

In order for this type of molecular shape matching to occur, the geometric descriptors of the surface patch must contain information related to both patch shape and position. The constructs of principal curvature and principal directions from classical differential geometry provide an elegant formalism that precisely captures these notions. The principal curvatures of the patch invariantly describe the shape of the patch, while the principal directions provide information related to its spatial orientation. Principal curvature and principal directions are informally described below; a more mathematically based explanation is given by Hosaka.[45]

The curvature, κ, measured at a fiducial point on a planar curve describes how the normal to the curve turns when moving along the curve. The curvature is negative for clockwise rotations and positive for anticlockwise rotations. Large magnitudes of κ indicate the curve is "more curved". When $\kappa = 0.0$, then the curve has no curvature and is simply a straight line. The principal curvature of a surface evaluated at a fiducial point on the surface is simply the curvature value of specific curves that are constrained to lie on the surface and pass through the selected fiducial point. A more formal description follows.

Let S denote a surface patch embedded in three-dimensional space (see Figure 1). Denote a reference point p that lies upon S, and let **n** be the unit normal vector to S at p. Construct a tangent plane t to S at p. Now construct a second plane containing **n** and orthogonal to t. Such a plane is called the normal plane and must intersect S. The intersection of the normal plane with S defines a plane curve termed a normal section. By construction the normal section is constrained to lie upon the surface S. The curvature κ at p of the normal section is termed the normal curvature at p. As the normal plane is rotated about the normal vector **n** the normal section changes and a new normal curvature may be calculated. After the plane is rotated by π radians, it returns to its starting orientation (although its normal points in the opposite direction). The minimum and maximum values of the normal curvatures (k_{min} and k_{max}) are the principal curvatures. The principal directions are the direction cosines of the vector defined by the intersection of the tangent plane t with the normal plane when the normal plane is aligned with the principal curvatures (see Figure 1). If the principal curvatures are equivalent to one another, then the point is termed an *umbilic* and depending upon the sign of the curvature, the local section of surface is either purely convex or concave. If the principal curvatures are distinct from each other, then the principal directions are orthogonal.[45] The fact that the principal directions are orthogonal will be used during the matching phase of the algorithm. The mathematical methods that we use for calculating these values will now be presented.

Shape Descriptor Calculation. Initially, the dot molecular surface of a molecule is computed.[46] Next, using the procedure defined by Lin et al.,[47] the critical points of the surface are computed. The critical points are defined as the center of gravity of each section (torus, spherical triangle, convex region)[48] of the molecular surface projected back onto the molecular surface (see Figure 2). To reduce the number of critical points used to describe a molecule, we do not use critical point locations associated with the toroidal sections of the surface. The critical points associated with either the spherical triangle or convex sections of the surface are the positions where the shape descriptor will be located. If, however, we consider the curvature at an infinitesimal area of surface, it will typically be spherical and not characteristic of a larger portion of the surface. Instead, we are interested in characterizing and describing a larger section of surface surrounding a surface point. Accordingly, we approximate a 2 Å radius section of surface surrounding a critical point by using a quadratic surface. The principal curvatures and directions are then calculated for the quadratically approximated surface section. This technique is similar to the method used by Zachmann et al.[49] for characterizing curvature domains of proteins.

It is important to realize that we are not interested in the strict mathematical definition of principal curvature and directions which relate to an infinitesimal area of a surface surrounding a reference point. Rather, we are characterizing and describing a larger section of surface. To distinguish the values that we calculate from their strict mathematical definitions, and the *global curvatures* calculated by Zachmann et al.,[49] we will henceforth refer to the values we calculate as the *local range curvatures* and *local range directions.* Because of the method used to calculate the local range curvatures and local range directions and the fact that these values are intimately associated with our shape

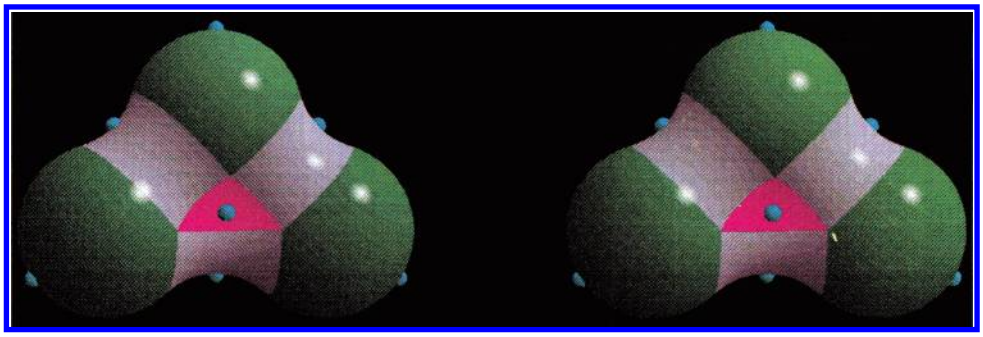

Figure 2. Molecular surface critical point locations as defined by Lin.[47] The critical point locations are denoted by small purple spheres. The green section is contact portion of the molecular surface. The spherical triangle and toroidal section of the reentrant surface are colored dark and light purple, respectively. The critical points associated with the toroidal sections of the surface are not used.

descriptor, we have named our descriptor the quadratic shape descriptor (QSD). The mathematical method for calculating a QSD will now be explained.

Formally, we denote the surface as the set of tuples $S = \{p_1, ..., p_n\}$, where $p = (v, n)$ is composed of the surface point location v in three-dimensional space and n is the unit vector normal to the surface at $p.v$. Similarly, the set of critical points is denoted as $C = \{c_1, ..., c_m\}$, where $c_i \in S$ and m is significantly less than n. For construction of the QSD associated with a critical point $c \in C$, we first determine a local patch of surface surrounding c. The surface neighborhood about critical point c is defined as

$$N(S, c, \rho) = \bigcup_{p \in S} \text{dist}(p.v, c.v) \leq \rho \tag{1}$$

The set $N(S, c, p)$ represents all of the surface points within a distance ρ from c. The distance measure we use is Euclidean distance through space and not a geodesic distance along the surface. The local range curvatures and directions are calculated for the section of surface represented by $N(S, c, p)$ using a least-squares fitting procedure.

The set $N(S, c, \rho)$ is transformed such that $c.v$ is located at the origin of the coordinate system with its normal $c.n$ aligned along the positive z axis. Letting (u, v, w) denote the coordinates of the transformed points in $N(q, \rho, S)$, a map relating the u, v coordinates to the displacement from the tangent plane to the surface at $c.v$ is constructed. Such a map is termed a "monge patch"[50] and has the form

$$x(u,v) = u\vec{\imath} + v\vec{\jmath} + w(u,v)\vec{k} \tag{2}$$

where $\vec{\imath}, \vec{\jmath}$, and $\vec{k}$ are orthogonal unit vectors constructed so that $\vec{k}$ is aligned to the normal $c.n$ and $\vec{\imath}$ and $\vec{\jmath}$ spans the plane tangent to the surface at $c.v$ (see Figure 3). The term $w(u, v)$ is known as the height function of u and v since it denotes the height of a point relative to the tangent plane as a function of the u and v coordinates. Recalling that $c.v$ is centered at the origin of the coordinate system with the normal aligned to the positive z axis, the local Hessian matrix (or the second fundamental form of the surface patch) can be expressed as[50]

$$\mathbf{II} = \begin{pmatrix} \partial^2 w/\partial u^2 & \partial^2 w/\partial u\, \partial v \\ \partial^2 w/\partial u\, \partial v & \partial^2 w/\partial v^2 \end{pmatrix} \tag{3}$$

The local range curvatures and directions of the surface patch are the eigenvalues and eigenvectors, respectively, of the **II** matrix.

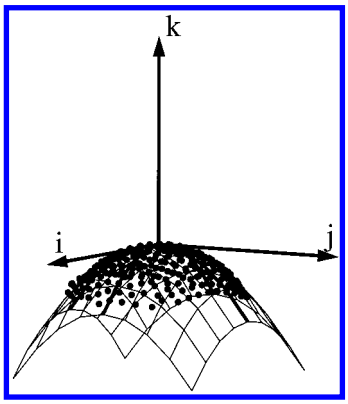

Figure 3. Transformed surface patch in preparation for construction of "monge patch". The critical point is at the center of the coordinate system, and the tangent plane is formed by the vectors $\vec{\imath}$ and $\vec{\jmath}$.

The terms of the matrix **II** can be calculated by fitting the transformed points in set $N(S, c, \rho)$ to the second-order part of the Taylor expansion of w:

$$\frac{1}{2}\left(\frac{\partial^2 w}{\partial u^2}u^2 + 2\frac{\partial^2 w}{\partial u \partial v}uv + \frac{\partial^2 w}{\partial v^2}v^2\right) \tag{4}$$

by using a least-squares fitting routine. Using matrix notation, and letting (u_i, v_i, w_i) denote the coordinate location of the ith point in the transformed set $N(S, c, \rho)$, the equation to fit can be expressed as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} \tag{5}$$

where

$$\boldsymbol{\beta} = \begin{bmatrix} \partial^2 w/\partial u^2 \\ \partial^2 w/\partial u\, \partial v \\ \partial^2 w/\partial v^2 \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} \frac{1}{2}u_1^2 & u_1 v_1 & \frac{1}{2}v_1^2 \\ \vdots & \vdots & \vdots \\ \frac{1}{2}u_n^2 & u_n v_n & \frac{1}{2}v_n^2 \end{bmatrix}, \quad \text{and}$$

$$\mathbf{y} = \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix} \tag{6}$$

The least-squares estimator of $\boldsymbol{\beta}$ is given by[51]

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{7}$$

The values of the partial derivatives, which are the elements of $\hat{\boldsymbol{\beta}}$, can then be substituted into eq 3 and used to calculate the eigenvalues and eigenvectors of **II**. These values are, as stated before, the local range curvatures and directions of the surface patch represented by $N(S, c, \rho)$. To place the

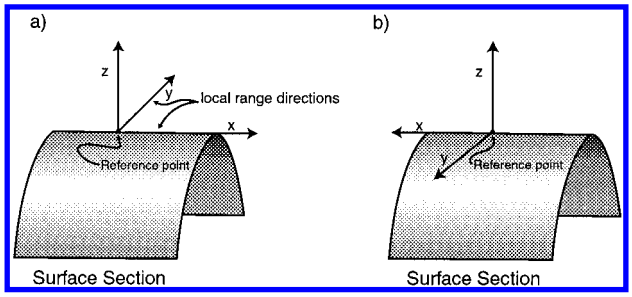

Figure 4. Two right-handed orthogonal coordinate systems constructed from the local range directions. The second coordinate system (b) is related to the first system (a) by a rotation of π radians about the *z*-axis.

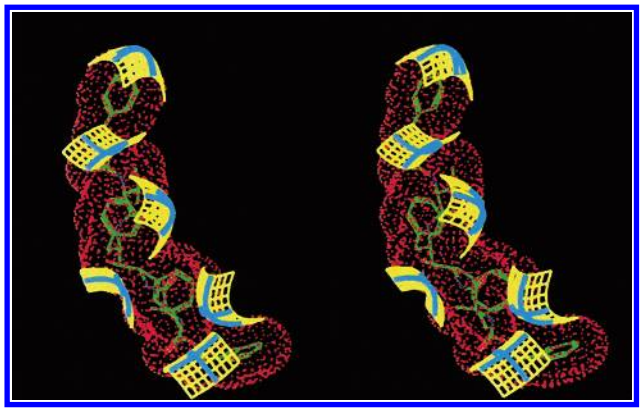

Figure 5. Molecular surface of A-74704 with six quadratic shape descriptors calculated at $\rho = 2.0$ Å. The molecular surface is in red, the surface nets which represent the quadratic shape descriptors are in yellow, and the local range curvatures in cyan. For visualization reasons, the molecular surface of A-74704 is displayed at 4 points/Å and not the actual density of 20 points/Å used in the calculation of the QSDs.

local range directions into the original frame of reference of the surface patch, they must be transformed by the inverse of the matrix used to position the set $N(S, c, \rho)$.

By letting (κ_{min}, λ_{min}) and (κ_{max}, λ_{max}) represent the local range curvatures and directions of the surface patch (where $\kappa_{min} \leq \kappa_{max}$ and λ_{min}, λ_{max} are the eigenvectors associated with the eigenvalues κ_{min} and κ_{max}, respectively), two right-handed orthogonal coordinate systems can be constructed (see Figure 4), the two being related by the 2-fold rotational symmetry about the normal to the surface at *c*. While the *x*-axis and *y*-axis of the coordinate system can be designated by λ_{min} and λ_{max}, respectively, the positive directions for these axes cannot be invariantly assigned. Therefore, the first coordinate system is created by using the normal to the surface as the positive *z* axis and then arbitrarily assigning one direction of λ_{min} as the positive *x* axis. The positive *y* axis can then be determined by the right-hand rule. The second coordinate system, which is related to the first by a rotation of π radians about the *z* axis, is not explicitly constructed but is treated implicitly during the recognition stage of the algorithm.

The parameter ρ in eq 1 determines how large a section of surface is approximated by a QSD. We have determined empirically that setting $\rho = 2.0$ Å captures the essential shape of small sections of molecular surface. Figure 5 displays the molecular surface of the HIV-1 protease inhibitor A-74704 along with six quadratic shape descriptors calculated at $\rho =$ 2.0 Å. The quadratic shape descriptors are displayed using a surface net. We display the QSD as a square patch for convenience and visual perception of local range directions, but the patch was fit only to a circular set of surface points. The local range directions are shown as thick curved lines on the net. Figure 5 demonstrates how the shape descriptor approximates the molecule surface but retains the salient local feature of the surface.

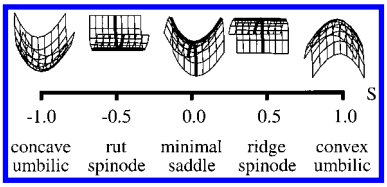

Figure 6. Continuum of quadratic shape represented by the shape index. The local range curvatures of the patches are shown as solid dark lines. The nets at $S = 1$ and $S = -1$ do not have local range directions shown since they are undefined for umbilic points. See text for details.

Shape Index. Aside from yielding the local range directions used to construct a QSD, the local range curvatures are rich in information that can be used to classify shape. One useful metric derived from the curvature values is the shape index *S*.[50] The shape index represents the degree of concavity or convexity of a local surface section and is defined by

$$S = -\frac{2}{\pi} \arctan \frac{k_{max} + k_{min}}{k_{max} - k_{min}} \qquad (8)$$

The shape index maps local shape at a particular location on the surface into the interval $[-1,1]$. All convex shapes have a positive index, while all concave shapes have a negative index. Figure 6 shows the continuum of shapes represented by the shape index *S*.

The shape index provides a convenient mechanism for determining the similarity between two sections of surfaces. The similarity measure for two surface patches with shape indexes S_1 and S_2 is

$$\mathrm{Sim}(S_1, S_2) = \frac{2.0 - |S_1 - S_2|}{2.0} \qquad (9)$$

Shapes that are identical will have a similarity score of 1.0, and shapes that are exactly opposite will receive a score of 0. The shape index can be used to to rank and/or prune potential matches during the matching phase of the algorithm as well as measure the extent of shape similarity between two surface patches. During the recognition stage of our algorithm, the similarity score is used to gauge the extent of surface similarity between two superposed molecular surfaces. We have reported elsewhere on studies using another shape classification called basic shapes.[37,38]

Surface Patch Matching. Let S_1 and S_2 represent two sections of surface characterized by their local range directions and curvatures calculated at fiducial points p_1 and p_2 (where p_1 lies on S_1 and p_2 lies on S_2). Consider the patch S_1. Let $\mathbf{k}_{min}$ and $\mathbf{k}_{max}$ denote the local range directions calculated at point p_1 on surface S_1 and let $\mathbf{n}$ represent the normal to the surface at p_1. A matrix $\mathbf{A}_1$, termed an alignment matrix, can then be constructed as

$$\mathbf{A}_1 = \begin{bmatrix} \mathbf{k}_{min} \\ \mathbf{k}_{max} \\ \mathbf{n} \end{bmatrix} \tag{10}$$

Since $\mathbf{k}_{min}$ and $\mathbf{k}_{max}$ are unit vectors that are orthogonal to each other as well as to $\mathbf{n}$, the alignment matrix $\mathbf{A}_1$ is an orthogonal matrix. A similar alignment matrix $\mathbf{A}_2$ can be constructed from the local range directions associated with surface patch S_2.

To superimpose surface patch S_2 onto patch S_1, the surface S_2 must be translated to the origin of the coordinate system, rotated so that its local range directions are aligned with those of S_1, and then translated so that fiducial point p_2 is superimposed on point p_1. The translations can be trivially calculated from the locations of the points p_1 and p_2. To align the local range directions of patch S_2 to those of S_1, the alignment matrix $\mathbf{A}_2$ must be rotated by a matrix $\mathbf{R}$ so that it becomes equivalent to the matrix $\mathbf{A}_1$. The desired rotation matrix can be calculated from simple matrix algebra.

$$\mathbf{A}_2\mathbf{R} = \mathbf{A}_1$$

$$\mathbf{R} = \mathbf{A}_2^{-1}\mathbf{A}_1 \tag{11}$$

The matrix $\mathbf{A}_2$ is an orthogonal matrix, and thus its inverse is equivalent to its transpose.[52] This fact can be used to simplify the computations necessary to calculate $\mathbf{R}$. Using homogeneous coordinates, the above rotation matrix $\mathbf{R}$ can be combined with the translation and the entire transformation represented as one 4×4 matrix. Let $\mathbf{T}_{in}$ represent the 4×4 translation matrix necessary to bring point p_2 to the origin of the coordinate system. Similarly, let $\mathbf{T}_{out}$ represent the translation matrix necessary to bring the point at the origin to the location specified by p_1. Finally, let the alignment matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ be represented as 4×4 rotation matrices. The matrix required to superimpose the two surface patches S_1 and S_2 is

$$\mathbf{R} = \mathbf{T}_{in}\mathbf{A}_2^{-1}\mathbf{A}_1\mathbf{T}_{out} \tag{12}$$

The above matrix can be calculated very quickly, requiring only three matrix multiplications.

Due to the 2-fold axial symmetry of the local quadratic shape descriptor (see Figure 6), the local range directions calculated at a point on a surface are determined up to a rotation of π radians about the normal to the surface at that point. Therefore, there are actually two rotation matrices that can be used to superimpose the two surface patches. The first matrix is calculated as outlined above. The second rotation matrix is calculated in the same manner except that the *direction* of the principal directions in the matrix $\mathbf{A}_2$ are negated prior to application of eq 12. The negation of the local range directions in matrix $\mathbf{A}_2$ is akin to rotating the alignment matrix $\mathbf{A}_2$ by π radians about the normal to the surface at point p_2.

The quadratic shape descriptor that we use to describe and orient local surface patches is defined as the tuple (c,S,$\mathbf{A}$). The tuple is composed of the location of the critical point c that lies upon the surface being described, the normal to the surface at c, and the local range curvatures and local range directions of the surface evaluated at c. The local range curvatures are represented in the form of the shape index S, and the local range directions and surface normal are contained within the alignment matrix $\mathbf{A}$.

Input:
- M Coordinates of Molecule
- ρ Distance parameter

Variables:
- $\mathbf{A}$ Alignment Matrix
- S Shape Index

Algorithm:

Create molecular surface for molecule M using the Connolly[46] algorithm.
Calculate critical points $C = \{c_1, \ldots, c_m\}$ of surface using Lin's[47] method.
for each $c \in C$
  $(c, S, \mathbf{A})$:= Create QSD at point c with distance range ρ
  Store $(c, S, \mathbf{A})$ to disk
end

Figure 7. QSD preprocessing algorithm.

QSD ALGORITHM

The algorithm that we have developed for shape-based similarity searching consists of two stages, an initial indexing or preprocessing section (which is done only once) and a run time object recognition phase. These two sections of the algorithm will now be explained.

Preprocessing. The preprocessing algorithm accepts as input the three-dimensional coordinates of a molecule and calculates the set of QSDs describing its surface shape. The pseudocode for the algorithm is listed in Figure 7.

The QSD preprocessing algorithm has a very low order of complexity. Ignoring the molecular surface and critical point calculations, the computation time and storage space complexity of the algorithm is *linear* in the number of critical points. This is in stark contrast to other molecular shape comparison algorithms that have either a cubic[53] or quadratic[54,55] complexity in the run time and storage requirements of their preprocessing phases.

For the molecular surface calculation we use a *united atom* model as Lin and Nussinov[56] have demonstrated that the explicit use of hydrogen atoms may be detrimental to molecular docking algorithms. The *united atom* van der Waals (VDW) radii are based upon values listed by Korolkovas.[57] A standard probe radius of 1.4 Å is used. Dot molecular surfaces are computed using Connolly's original molecular surface algorithm.[46,58] The placement of points upon the molecular surface of a molecule by this algorithm depends slightly upon the orientation of the molecule. The uncertainty in surface point location induces uncertainty in the critical point locations, critical point normal orientations, and the local range curvatures and directions. To ameliorate the effects this non-invariant placement of surface points has upon our QSD calculation, we compute the surface at a high density of 20 dots/Å^2. Such a high density was selected for several reasons. First, Lin et al.[47] has shown that the uncertainty in critical point location and surface normal orientation is reduced when densely distributed surface points are used for their calculation. (At a density of 10 points/Å, Lin et al. calculate a worst case error of 0.15 Å in critical point location and 5° for normal orientation. Although we did not determine the error associated with critical point location and normals at 20 points/Å, it should be below the values calculated at the density of 10 points/Å.) Zachmann et al.[49] have shown, by using surface points distributed on a sphere, that the numerical calculation of curvature converges toward its analytical value as surface point density is increased.

Input:

M_T, M_M	Coordinates of Target Molecule and Moving Molecule
Q_T, Q_M	QSD set describing Target Molecule and Moving Molecule
ΔS	Shape Filter

Variables:

q_t, q_m, q', q_t'	Single QSD
ML	match list, used to stores pairs of matching QSDs
G	A cubic grid of 1Å resolution
	/* The grid encompasses QSD set Q_T */

Algorithm:

```
Impress QSD set Q_T into grid G
for each q_t ∈ Q_T
    for each q_m ∈ Q_M
        if (|q_t.S − q_m.S|) ≤ ΔS)
          for i = 1 to 2
              Using Equation 12 calculate transformation matrix R
              /* Matrix R superposes q_m onto q_t */
              if (i == 2 )  /* Second time through */
                 Rotate R by π radians about normal to critical point q_m.c
              endif
              ML := ∅
              for each q ∈ Q_M
                  q' := q transformed by matrix R
                  Using Grid G match q' to closest q_t' ∈ Q_T
                  Update matchlist ML with pair (q', q_t')
              endfor
              Using Equation 13 score surface similarity of ML
              if (sufficient similarity)
                    store alignment
              endif
          endfor
        endif
    endfor
endfor
```

Figure 8. QSD matching algorithm.

Recognition. The superimposition stage of the algorithm uses the precomputed shape information to quickly align two molecules. The algorithm is listed in Figure 8. One molecule which remains stationary during the matching is termed the target molecule, while the other molecule is designated the mover. First, a cubic grid with voxel size of 1.0 Å is constructed and critical points associated with the target molecule are placed into the grid. A critical point is placed into a voxel if it is within 1.0 Å of the voxel center. Thus, each critical point can appear in the grid more than once. The placement of the critical points into the grid in this fashion facilitates nearest neighbor searches when the molecular alignments are generated.

Alignments are generated by using eq 12 to compute the matrix neccessary to superimpose two surface patches (one from each molecule). This matrix is then used to transform all of the QSDs on the moving molecule into the frame or reference of the target molecule. Using the cubic grid, the closest target critical point within 1.0 Å to each of the newly transformed moving critical points is determined. A pairing or match list of target to mover critical points is then constructed. This pair list is a bijection, namely, each target critical point in the match list can match one and only one moving critical point, and each moving critical point in the list is matched by at most one target critical point. The match list takes the form $ML = \{ml_0, ..., ml_n\}$, where $ml = (t_i, m_j)$, indicating that the ith target critical point matches the jth moving critical point. The criterion for a match is that the distance between t_i and m_j must be less than 1.0 Å. If more than one critical point on the moving molecule is within 1 Å of a critical point on the target molecules, a match is formed between the target and moving critical points that are separated by the smallest distance. We have found that recomputing the transformation matrix to fit all the points in the match list in a least-squares sense does not appreciably change the quality of the results; therefore, we do not recompute the alignment matrix. The alignment is then scored using the shape index. Let $S(\text{ml}.x)$ represent the value of the shape index S for the match list QSD ml.x. Based upon eq 9, the surface similarity score, Y, for a matching is defined as

$$\mathrm{Y} = \sum_{\text{ml}\in\text{ML}} \text{Sim}(S(\text{ml}\cdot t_i), S(\text{ml}\cdot m_j)) \qquad (13)$$

The surface similarity score is used to collect high-scoring alignments. Typically, we save the top 100 alignments as ranked by Y. These alignments are then further scored using the more time consuming common volume metric.[59,60] Another metric based only on shape statistics has also been used and is reported elsewhere.[61]

As stated earlier, because of the C_2 symmetric nature of a QSD, there are actually two transformation matrices that can be used to superimpose a pair of QSDs. Therefore, the recognition procedure must be repeated for the C_2 symmetric matching of the two QSDs. Note, however, that while the QSD has a 2-fold axial symmetry, the entire molecule generally does not share that symmetry. Therefore, the two equivalent orientations of the QSDs result in two nonequivalent molecular orientations (see Figure 9).

The common volume metric is calculated using a grid based algorithm with a resolution of 1.0 Å^3. The target molecule along with the aligned mover molecule are placed into the grid. Voxel centers of the grid that are within the van der Waals radius of an atom on both the target and mover molecules are turned on. The common volume score is then the count of the number of voxels that are on.

Shape Filtering. During the generation of alignments, the shape index may be used to prune the search space. If the target molecule contains m critical points and the mover n critical points, then $2mn$ transforms need be investigated. Using the heuristic that similarly shaped sections of surface will lead to good alignment, we prune potential matches by only constructing transformation matrices if the two surface patches represented by the two critical points are "similar" in shape. We define a parameter called ΔS that is used for shape filtering. Shape filtering is not used to improve the quality of the results but rather to reduce the number of transformation matrices constructed. If shape filtering is enabled during the matching, then transformation matrices are constructed and alignments generated only for pairs of QSDs for which the absolute value of the difference in their shape index is less than or equal to ΔS.

(a) Self-Shape Matching. The effects of shape filtering can be demonstrated by performing a self-matching test. The self-matching test consists of selecting a base location for a ligand (L1), randomly rotating and translating a duplicate of the ligand (L_2), computing the molecular surface and shape descriptors for each position of the ligand, and then attempting to superimpose L_2 onto L_1. Table 1 demonstrates the effects of shape filtering when the algorithm is used for the self-matching of the HIV-I protease inhibitor A-74704. For *identity* matching, the transformation matrix necessary to superimpose one orientation of the ligand onto the other can be identified even when the value of ΔS is set quite low. This is because the ligands L_1 and L_2 fit perfectly and the

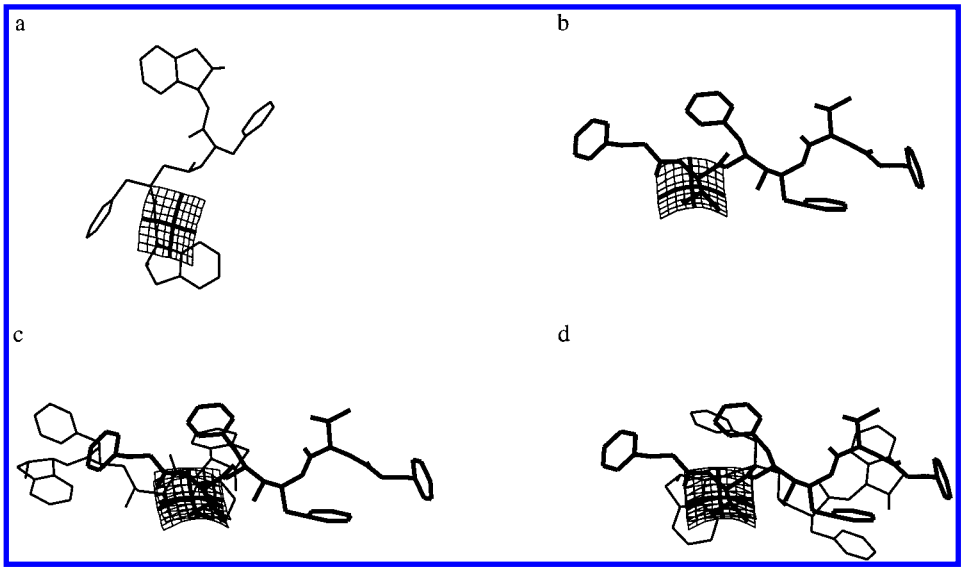

Figure 9. Matching algorithm steps. (a) Selection of a QSD on the moving molecule. (b) Selection of a QSD on the target molecule. (c) One superimposition of the QSDs from the target and moving molecule. The superposition of the molecules generated by the QSD superimposition is quite poor and would be rejected by the scoring algorithm. (d) The C_2 symmetric superposition of the same QSDs as in c. The moving molecule now superposes very nicely on the target molecule.

Table 1. Ligand A-74704 Matched to Randomly Positioned Self

ΔS	best RMS (Å)	% red[b]	CPU[c] (s)
∞	0.04	0	48.82
0.50	0.04	54	25.04
0.33	0.04	66	18.72
0.10	0.04	88	9.23
0.05	0.04	92	6.32
0.01	0.20	98	3.58

[a] RMS of the heavy atoms of A-74704 after positioning. [b] The percentage of transforms that were not used due to shape filtering.[c] The amount of CPU time in seconds required to complete the matching run. The CPU time relates to a DEC Alpha 3000/400 under OpenVMS.

algorithm is able to locate nearly identical sections of surface and then superimpose the ligands. However, for *similarity* searching, if ΔS is set too small, then the correct alignment may be overlooked. This is because similar and not identical surface patches need to be matched. Although ΔS can be set as low as 0.1 and realistic molecular alignments can still be generated, we have found empirically that setting $\Delta S = 0.3$ eliminates approximately two-thirds of the transformation matrices without the risk of missing reasonable and interesting molecular alignments.

For the self-matching test, the deviations of the heavy atoms for the best superimpositioning of the ligand to itself is 0.04 Å and not exactly zero as one would expect for a perfect match. The reason for this small error is that the molecular surface and shape descriptors are generated twice, once for each location of the ligand. As mentioned earlier, the location of the molecular surface points as generated by the original Connolly molecular surface algorithm[46,48,58] are not invariantly located on the surface of the molecule; they depend slightly on the orientation of the molecule. The small deviations in the locations of the molecular surface points cause small deviations in the values calculated for the shape descriptors. These small differences in the values of the shape descriptors cause imperfect self-alignments. To gauge the effects that the non-invariant placement of molecular surface points has on our algorithm, we repeated the self-matching test for A-74704 20 times. For each run we set $\Delta S = 1.0$.

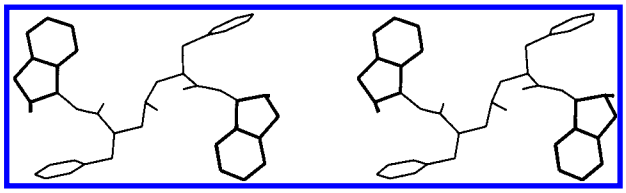

Figure 10. Self-subshape matching. Two superimpositions determined by QSD of the hydroxy−indanyl fragment of L-700,417 positioned onto the same ligand. The superposed hydroxy−indanyl ring fragments are shown as thick dark lines. See text for explanation.

The average root mean square deviations in the distance of the heavy atoms of the superimposed molecules is 0.10 Å with a standard deviation of 0.05 Å. These values indicate that the algorithm is not overly sensitive to the generation of the molecular surface. Because of the deterministic nature of the algorithm and the method used for matching, our technique will always perfectly superimpose two identical molecules. This is in contrast to stochastic algorithms which do not guarantee that an optimal superimposition of identical rigid molecules will be found.[10]

(b) Self-Subshape Matching. We also tested the subshape matching performance of our algorithm. In this test we used the ligand L-700,417 from the Brookhaven Protein Data Bank (PDB)[62] file 4phv. This test consists of selecting the entire ligand as the target molecule and a fragment of the molecule as the mover molecule. The fragment we selected is one of the two hydroxy−indanyl ring systems. The locations of both the molecule and the fragment were randomized and the molecular surfaces and quadratic shape descriptors calculated. We then used our algorithm with $\Delta S = 0.3$ to superimpose the fragment back onto the target molecule. The algorithm was successful in superimposing the fragment onto both of the hydroxy−indanyl ring systems present in the ligand. The root mean square (RMS) deviations in positioning for the fragment is 0.05 and 0.06 Å (see Figure 10). The entire matching process required 1.95 s of CPU time on a DEC-ALPHA 3000/400.

Complexity. The following analysis of the computational complexity of the QSD recognition algorithm is based upon

Table 2. Molecules Matched and Preprocessing Times

ligand	PDB	resolution (Å)	active site residues	RMS[a] (Å)	CPU[b] (s)
L-700,417	4PHV	2.1	residue 25−27 in chains A and B	0.15	39
MVT-101	4HVP	2.3		0.24	50
A-74704	9HVP	2.8		−	54
NAPAP[c]	lETS	2.3	189−195, 214−220, 57, 102	0.26	35
4-TAPAP[d]	1ETT	2.5		0.34	30
PPACK[e]	1DWE	3.0		0.44	25
MQPA[f]	1ETR	2.2		−	33
DFKi[g]	4EST	1.8	41, 57, 102, 189−195, 213−216, 226−228	0.34	33
TOMI[h]	IPPF	1.8		−	384

[a] Root mean square deviation in atomic position of active site backbone atoms after least-squares fitting of the mover and target proteins. In each case a "−" indicates the target protein. For example, the RMS when 4PHV is aligned to 9HVP is 0.15 Å. [b] CPU time in seconds for preprocessing which includes molecular surface calculation and generation of the shape descriptors. [c] 4-TAPAP: N^{α}-(4-toluenesulfonyl)-DL-*m*-amidinophenylalanylpiperidine. [d] NAPAP: N^{α}-(2-naphthylsulfonylglycyl)-DL-*m*-amidinophenylalanylpiperidine. [e] PPACK: D-Phe-Pro-Arg-chloromethylketone. [f] MQPA: (2*R*,4*R*)-4-methyl-1-[N^{α}3-methyl-l,2,3,4-tetrahydro-8-quinolinesulfonyl-L-arginyl]-2-piperidine carboxylic acid. [g] DFKi: Ac-Ala-Pro-Val-difluoro-*N*-phenylethylacetamide. [h] TOMI: third domain of turkey ovomucoid inhibitor.

the listing in Figure 8. Let n represent the number of QSDs used to describe the shape of the target molecule, and let m be the number on the moving molecule. In the two outermost loops of the algorithm each target QSD is matched to each moving QSD. For each of these pairings two transformation matrices are calculated. Therefore, the total number of the matrices calculated is $2mn = O(mn)$. Note that the constant value of 2 does not affect the asymptotic behavior of the algorithm. For each matrix calculated, all of the QSDs in the moving set are transformed, matched with QSDs in the target set and then the surface similarity score assessed. This procedure, which is analogous to the voting process in the geometric hashing algorithm of Nussinvo and Wolfson,[53] is linear in the number of transformed points. The total complexity of the algorithm is thus $nm \times m = O(nm^2)$. The bound of $O((nm^2)$ is a worst case bound. Typically, shape filtering is used to reduce the total number of matrices and thus the number of molecular alignments that are actually computed (see Results Section, Table 4).

EXPERIMENTAL SECTION

The algorithm presented above is implemented in a computer program called QSD (quadratic shape descriptor). The program is written in the C programming language.

To demonstrate the utility of our algorithm, we present the matching of inhibitors from three different protein systems: HIV-1 protease, thrombin, and serine protease. In all cases the molecules matched were taken from the Brookhaven Protein Data Bank.[62] For each test case we have performed three different matchings, one with $\Delta S = \infty$ and the others with $\Delta S = 0.3$ and $\Delta S = 0.1$. All execution times are given in CPU seconds and relate to a DEC alpha 3000/400.

Experimental alignments for ligands that inhibit the same protein system are generated by the least-squares fitting of the active site backbone atoms of the protein cocrystallized with the ligand. The same transformation matrix is then applied to the ligands. These "experimental alignments" are indicative of how *nature* views the molecules as being similar and therefore provide an unbiased metric to measure the success of the algorithm. Due to the resolution of the crystal structures, these experimental alignments are subject to error. The average positional errors of atoms in a crystal structure is approximately one-sixth the resolution of the crystal structure.[63] The average resolution of the crystal structures used in this study is 2.31 Å, and thus the average uncertainty in the locations of the molecules used in this study is approximately 0.39 Å. The experimental alignments that we use for measuring the success of our algorithm are generated by superimposing two crystal structures, and thus, the experimental errors associated with each structure must be propagated into the final result. The final error associated with these experimental alignments is approximately 2×0.39 Å $= 0.78$ Å. This value can be used when assessing the quality of the results generated by the QSD superimposition algorithm. Tables 2 and 3 list the molecules matched as well as other pertinent information. The structures of the molecules used in this report are shown in Figure 11.

Table 3. Crystallographic Scores

target		mover			
ligand	no. of descriptors	ligand	no. of descriptors	vol overlap	Y
A-74704	154	L-700,417	152	250	45
A-74704	154	MVT-101	145	297	43
MQPA	106	NAPAP	116	271	29
MQPA	106	4-TAPAP	92	266	35
MQPA	106	PPACK	87	179	22
TOMI	904	DFKI	110	274	38

RESULTS

HIV-1 Protease Inhibitors. For the matchings of the HIV-1 protease inhibitors the ligands MVT-101 and L-700,417 were both individually aligned to the ligand A-74704. In all cases examined the algorithm finds superpositions of the ligands consistent with the experimental alignment (see Table 4). When compared to the experimentally determined alignments, the best superimpositions found have an RMS difference in the atomic positions of less than 1.00 Å. Ranking the superimpositions according to either volume overlap or surface similarity score selects alignments with low RMS values. The best alignments found for L-700,417 to A-74704 is the C_2 symmetric orientation of the ligand (we have adjusted the reported RMS value accordingly). Furthermore, the amount of CPU time required to compute all the alignments is very small. Figure 12 shows the best matching of L-700,417 to A-74704, and Figure 13 shows the best matching of MVT-101 to A-74704. In both of these figures the value of $\Delta S = 0.3$ and the superimpositions with

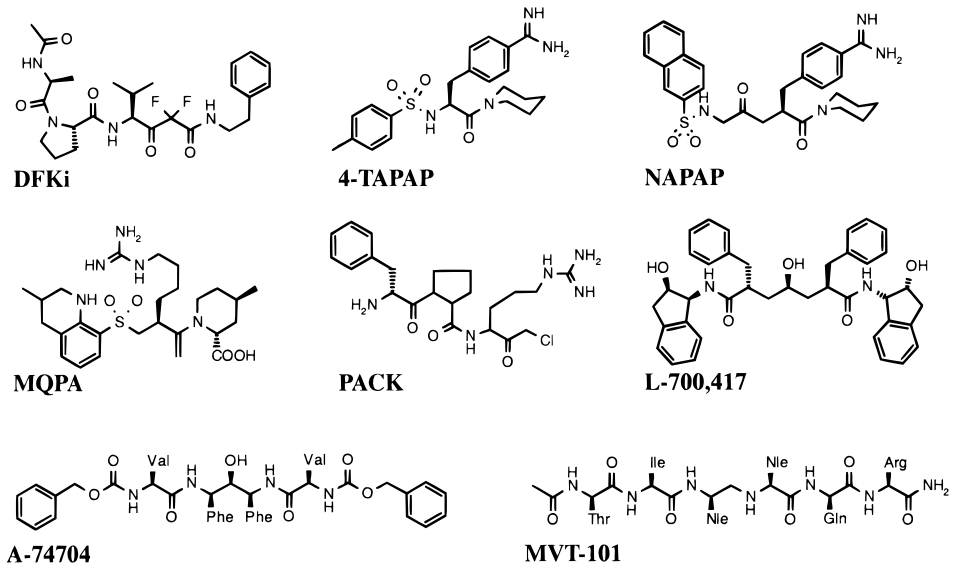

Figure 11. Structures of the molecules used to test the QSD matching algorithm.

Table 4. QSD Matching Results

		best RMS[a] (Å)					best vol[b]		best Y[c]		
target	mover	ΔS[e]	% red[f]	RMS (Å)	vol	Y	RMS (Å)	vol	RMS (A)	Y	CPU[d]
A-74704	L-700,417	∞	0	0.47	251(1)	99(1)	0.47	251	0.47	50	54.57
A-74704	L-700,417	0.3	69	0.47	251(1)	99(1)	0.47	251	0.47	50	19.01
A-74704	L-700,417	0.1	87	0.89	235(3)	75(3)	1.48	246	1.21	43	8.51
A-74704	MVT-101	∞	0	0.92	262(3)	73(4)	1.18	273	1.18	44	42.99
A-74704	MVT-101	0.3	70	0.92	262(1)	73(1)	0.92	262	0.92	37	14.52
A-74704	MVT-101	0.1	90	0.92	262(1)	73(1)	0.92	262	0.92	37	7.64
MQPA	NAPAP	∞	0	0.72	275(1)	58(15)	0.72	275	2.14	39	22.27
MQPA	NAPAP	0.3	71	0.72	275(1)	58(11)	0.72	275	2.14	39	8.90
MQPA	NAPAP	0.1	89	0.72	275(1)	58(8)	0.72	275	2.37	38	5.38
MQPA	4-TAPAP	∞	0	0.83	265(1)	71(1)	0.83	265	0.83	35	16.31
MQPA	4-TAPAP	0.3	71	1.08	260(1)	68(5)	1.08	260	1.13	34	7.71
MQPA	4-TAPAP	0.1	89	1.10	246(4)	61(4)	1.18	254	2.81	32	4.36
MQPA	PPACK	∞	0	0.68	190(1)	49(5)	0.68	190	8.80	27	14.03
MQPA	PPACK	0.3	71	0.68	190(1)	49(3)	0.68	190	2.24	27	5.41
MQPA	PPACK	0.1	90	0.68	190(1)	49(2)	0.68	190	8.76	25	3.30
TOMI	DFKI	∞	0	0.94	264(2)	59(6)	1.13	273	22.82	33	173.37
TOMI	DFKI	0.3	69	0.96	264(1)	57(5)	0.96	264	10.72	31	66.71
TOMI	DFKI	0.1	89	1.46	245(3)	55(8)	1.53	249	16.26	32	24.53

[a] The superimposition with the lowest RMS value to the experimentally determined orientation. The RMS value, the volume overlap score, and the surface similarity score for the superimposition are listed in the fifth, sixth, and seventh columns of the table. Numbers in parentheses are the rankings of the orientation according to the column headings. [b] The superimposition with the highest volume overlap score. The RMS value for this superimposition to the experimentally determined orientation is listed in column eight. The volume overlap is listed in the ninth column. [c] The superimposition with the highest surface similarity score. [d] CPU time in seconds on a DEC-ALPHA 3000/400 required to complete and score the matching. [e] ΔS is the value of the shape filter (see text). [f] The percentage of transforms eliminated by using the shape filter at the indicated value.

the lowest RMS values are also the superimpositions with the highest volume overlap and surface similarity scores. Surface similarity scoring here works very well.

Table 4 also clearly demonstrates the effects of shape filtering. When ΔS is reduced from ∞ to 0.3, the run time of the algorithm is substantially reduced and yet the quality of the results is not affected. When shape filtering is employed, a large percentage of the transformation matrices are never constructed. This is due to the fact that the surface patches (one from each molecule) leading to these superpositions are not sufficiently similar to each other and thus not used to create a superimposition of the two molecules. For example, when L-700,417 is matched to A-74704, and $\Delta S = 0.3$, approximately 70% of all possible alignments are never constructed. The ligand A-74704 has 152 QSDs on its surface, while the ligand L-700,417 has 145 (see Table 3). When shape filtering is not used (i.e., $\Delta S = \infty$), the total number of potential superpositions is $2 \times (145 \times 152) = 44\,080$. However, when $\Delta S = 0.3$, then 13 224 superpositions are investigated, and when $\Delta S = 0.1$, then only 10% or 4408 superpositions of the two ligands are evaluated. It is remarkable that so few orientations of the two ligands are investigated, yet realistic molecular superimpositions can be located. The reason why the quality of the results is not drastically affected when shape filtering is used is that the majority of alignments generated from dissimilarly shaped surface patches do not lead to reasonable molecular superimpositions. Such a result underscores the ability of the local quadratic shape descriptor to capture and represent molecular surface shapes.

Thrombin Inhibitors. The three Thrombin inhibitors 4-TAPAP, NAPAP, and PPACK were all individually aligned to the target molecule MQPA. Again, in all cases the best alignments found are consistent with the experi-

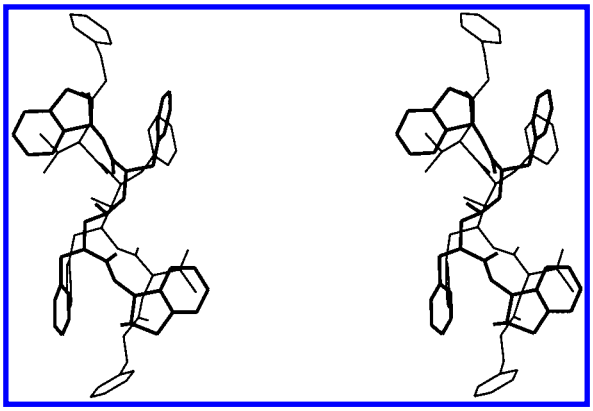

Figure 12. Stereoview of the best matching of L-700,417 (thick lines) onto A-74704 (thin lines) found when $\Delta S = 0.3$.

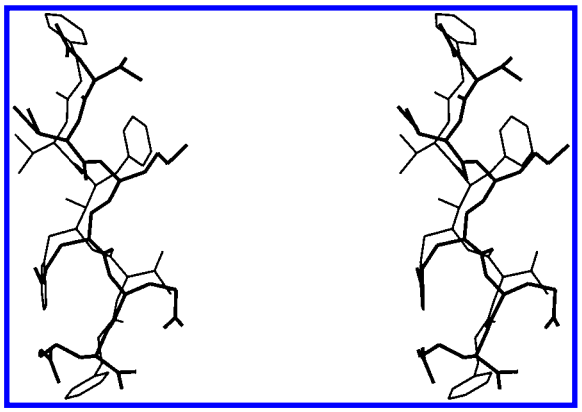

Figure 13. Stereoview of the best matching of MVT-101 (thick lines) onto A-74704 (thin lines) found when $\Delta S = 0.3$.

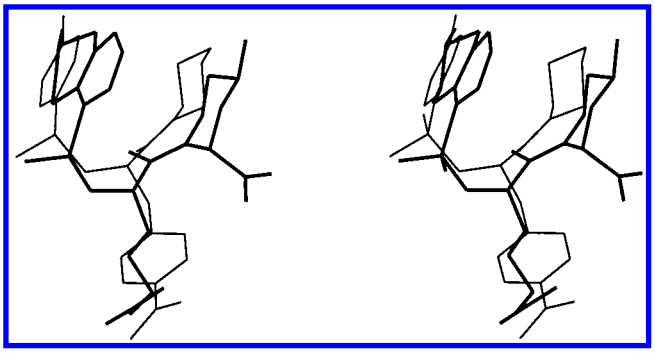

Figure 14. Stereoview of the superimposition of 4-TAPAP (thin lines) onto MQPA (thick lines) with the highest volume overlap score as generated by QSD with $\Delta S = 0.3$.

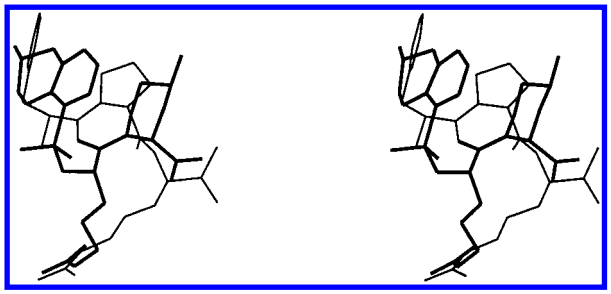

Figure 15. Stereoview of the superimposition of PPACK (thin lines) onto MQPA (thick lines) with the highest volume overlap score as generated by QSD with $\Delta S = 0.3$.

mental superpositions (see Table 4). Figures 14−16 show the best alignments selected on the basis of the common volume of 4-TAPAP onto MQPA, NAPAP onto MQPA, and PPACK onto MQPA. These alignments were generated with

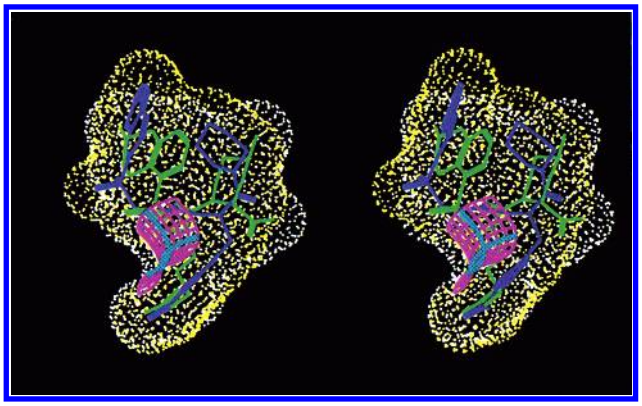

Figure 16. Stereoview of the superimposition of NAPAP onto MQPA with the highest volume overlap score generated by QSD with $\Delta S = 0.3$. MQPA is colored green and its molecular surface is white. NAPAP is colored blue and its molecular surface is yellow.

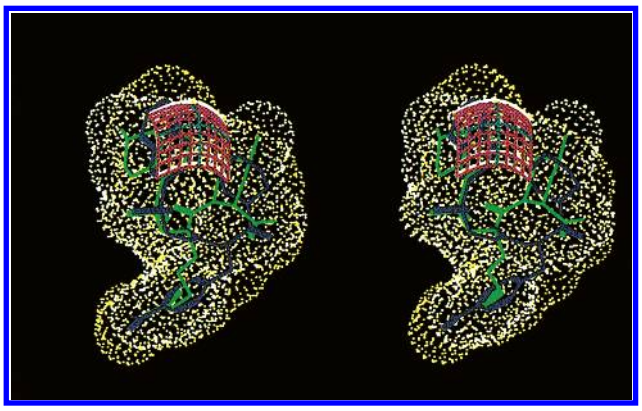

Figure 17. Stereoview of the superimposition of NAPAP onto MQPA with the highest surface similarity score generated by QSD with $\Delta S = 0.3$. Same color scheme as in Figure 16.

$\Delta S = 0.3$, and in all cases the alignments with the highest common volume are the superimpositions closest to the experimentally determined alignments.

Figure 16 displays the molecular surfaces of the two ligands MQPA and NAPAP as well as the quadratic shape descriptors used to generate the alignment. (For visualization clarity the surfaces are shown with a point density of 4 points/Å and not the actual density of 20 points/Å used in the calculations.) The value of the shape index S for the two descriptors shown are 0.22 and 0.23 for MQPA and NAPAP, respectively. Because the descriptors are so similar in shape, they perfectly superimpose and thus only one descriptor is visible in the image. Figure 17 shows the superimposition of NAPAP onto MQPA with the highest surface similarity score. The RMS deviation of this orientation of NAPAP to the experimental orientation is 2.14 Å. Although the RMS value is high, this superimposition seems reasonable and is actually ranked third by the common volume score.

Serine Protease Inhibitors. The serine protease inhibitors provide an interesting challenge to our algorithm. The size of the inhibitors differs considerably; TOMI is a 56 residue peptide inhibitor of Turkey ovomucoid, and DFKi is a tripeptide inhibitor of elastase class proteins. This example demonstrates the ability of QSD to find subshape similarity and also underscores the usefulness of shape filtering. For this example TOMI was designated as the target molecule and DFKi as the mover. When no shape filtering is used, the algorithm finds an alignment of DFKi onto TOMI with

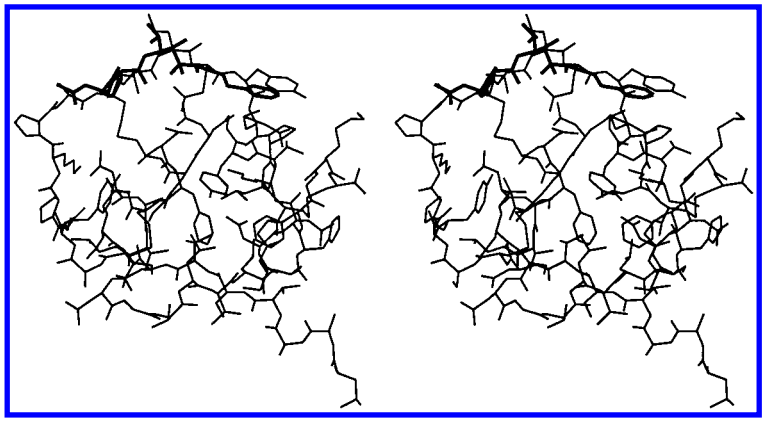

Figure 18. Best superimposition of DFKi (thick lines) onto TOMI (thin lines) selected on basis of volume overlap. This superimposition was generated with $\Delta S = 0.3$.

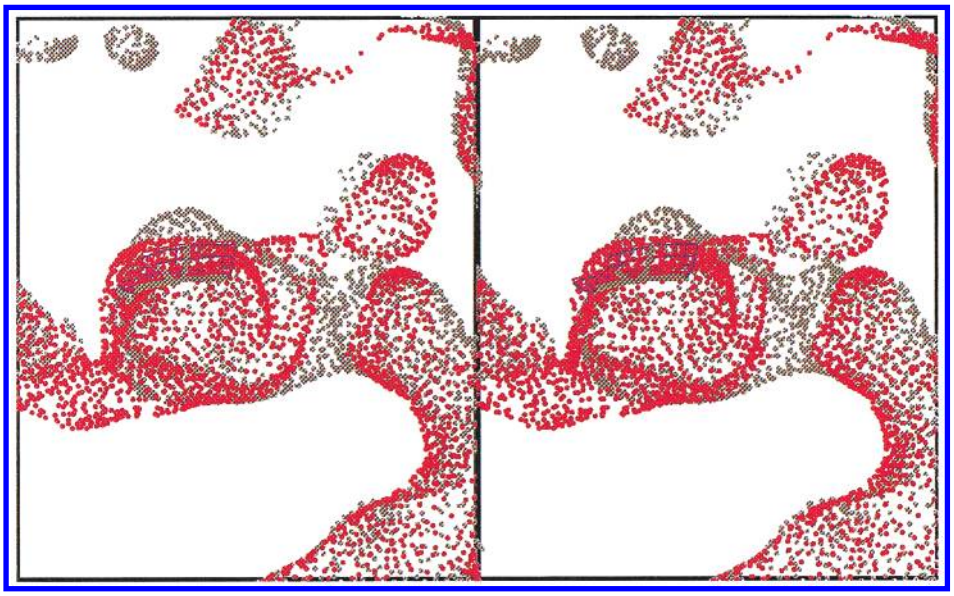

Figure 19. Superposition of sheep COX-1 (lpge, red) and mouse COX-2 (4cox, gray) active sites. Aligned quadratic shape descriptors on the top of site have the same shape. The view is looking into the site from the channel. These two proteins have 68% homology, but the active sites are similar. The COX-2 site is larger.

an RMS of 0.94 Å from its crystal orientation in approximately 173 s of CPU time (see Figure 18). When the shape filtering is used at a level of $\Delta S = 0.3$, the same alignment is found in 66.7 s of CPU time. When $\Delta S = 0.1$, the run time of the algorithm is further decreased to 22.53 CPU s; however, the quality of the results deteriorates slightly (RMS = 1.46 Å).

Alignment of Protein Active Sites. We were also interested in the reverse of aligning ligands of proteins, namely, aligning the proteins themselves on the basis of the shape of their active sites. If the active sites are conserved in shape within a family of different proteins, then one might expect to be able to align the proteins so their active sites are superimposed. Typically one aligns such proteins by the α carbons of the conserved residues in the active site. Aligning just by shape is more general. To test this application, we selected the cyclooxygenase family of COX-1 and COX-2. From the PDB the pgel structure[64] for sheep COX-1 and the 4cox structure[65] were selected (the A chain), and ligands and water were removed. Shape descriptors were derived for the entire surface of the proteins, and the QSD algorithm found the best superpositions. For higher efficiency one can carve around the active sites or only use QSD's on the active site surface. The fact that one can align active sites without biasing it toward the active site is interesting. Figure 19 shows the two active sites aligned, with the two QSD's which have the same shape value. The COX-1 site (red) is smaller than the COX-2 site (gray), but this QSD corresponding to the surface region of alanine-527 gives a good alignment of 0.762 root mean square deviation based on Arg-120, Ser-530, and Tyr-385 α carbons. The resolution of the 1pge crystal structure is 3.5 Å, and for 4cox it is 3.0 Å. The top six alignments based on shape similarity or common volume gave excellent alignments of the active sites, with the highest RMSD of 1.261 Å.

DISCUSSION

The QSD algorithm avoids the combinatorial explosion inherent in molecular matching by extracting and using the information that exists in the geometry of the molecular surfaces of the molecules being matched. The extracted information, in the form of local range curvatures and directions of surface patches, provides a mechanism for the superposition of two dissimilar molecules. The QSD superpositioning algorithm examines the surface of two distinct molecules for small regions that are similar in shape. The two molecules are then brought into a registration as dictated by the alignment of these small similarly shaped surface

sections. The alignment of local surface patches provides a powerful and convenient mechanism for generating globally reasonable molecular superpositions. Furthermore, through the use of shape filtering, the QSD algorithm efficiently reduces the search space for superposition-based similarity searching.

It is important to realize that our method is not an optimization procedure in that the alignment is not adjusted to optimize an objective function. Rather, the QSD algorithm intelligently places molecules in relation to each other and then scores the resulting alignment. It is interesting to note, however, that the best alignments found often have a volume overlap exceeding the value for the crystal alignment. The ligands in the crystal interact with the receptor as a result of a variety of chemical interactions (e.g., electrostatic, hydrophobic, hydrogen-bonding). Since our matching is based exclusively on shape, it is surprising that our algorithm finds the crystallographically determined position. The fact that it does is evidence of the importance of shape alone for determining molecular alignments.

The generality of the method is shown by alignment of ligands and by aligning cavities in proteins. The fact that it found common shape similarity between the 56 residue TOMI and the tripeptide DFKi illustrates its ability to find common subshapes between objects quite different in size. Similarly, it would find similarly shaped cavities within proteins of different structure and size.

In analogy with Perkin's method,[28] a strength of our algorithm is that we provide the user with a series of realistic alignments based upon surface and steric similarity. While it is generally accepted that molecules with similar shape perform similar functions, it should be realized that steric similarity is a necessary, but not sufficient, condition for molecules to exhibit the same biological effect. Chemical similarity should be taken into account. In each of the experiments we have presented in this paper, our algorithm finds superpositions of molecules that are extremely close to the experimental alignment. Selection of the alignment with the largest common volume or surface similarity score is not always the superposition closest to the crystal structure. Alignments that are close to the crystal structure are always within the top 15 superpositions as ranked by surface similarity, and within the top 10 when ranked by volume. Thus, a short list of sterically feasible alignments can be constructed and more sophisticated similarity metrics that include chemical information can be used to provide an alternative ordering of the alignments.

A major strength of the quadratic shape descriptor is that it is not limited to describing *knobs* and *holes*. The quadratic shape descriptor that we have introduced may represent any surface shape that can be approximated by a quadratic function. This feature makes our descriptor very versatile since it can describe a plethora of shapes. A significant drawback of our approach is that we must match rigid molecules since the conformational flexibility of the ligands is not addressed during the matching process. However, the QSD algorithm is sufficiently fast that it is feasible to incorporate ligand flexibility by performing multiple matches on a set of low-energy conformers of the molecules being matched.

Another application for the quadratic shape descriptors introduced in this paper is for shape-based molecular fingerprinting. Several similarity searching algorithms use counts of molecular fragment occurrence or histograms of interatomic distances as a means of describing molecular shape.[21−24] These discrete counts are then used to index databases of molecular structures and to determine the extent of similarity between pairs of molecules. A similar technique can be accomplished using the quadratic shape descriptors. For each molecule in the database, the frequency distribution of shape index values that exist on the surface of that molecule can be calculated. Molecules that are globally similar in shape should have shape index distributions that are also similar. Thus, the similarity between two molecules can then be assessed by comparing the shape index distributions. Work in this area is underway.

A satisfying property of our similarity searching algorithm is that it can be easily adapted to search for surface complementarity. Since shape complementarity is fundamental to ligand−receptor binding,[66] the QSD algorithm can be used for predicting the bound orientation of a ligand into its receptor protein. The self-matching test discussed earlier in this report can be considered a form of ligand docking. In this case, one copy of the molecule is designated the ligand and the inverted surface of the other copy is viewed as a perfectly shaped complementary active site. The ligand is then "docked" into this site. For molecular docking our recognition algorithm is unchanged except for the preprocessing phase. Ligand molecules are preprocessed in the same manner, but, for complementarity searching, the receptor molecule's surface is inverted. The procedure for inversion of the surface is simply to change the direction of the normals to the surface so that they point into the receptor. The application of our method for the docking of ligands to their receptor proteins[39] and for measuring shape complementarity[37,38] will be presented elsewhere.

The use of the Connolly dot surface in conjunction with the critical points as defined by Lin et al.[47] provides a convenient mechanism for calculation and location of the quadratic shape descriptors upon the molecule surface of the molecules to be matched. However, the raw Connolly surface may not be the best surface to use, because the salient geometric features are difficult to identify amid the small bumps and grooves of the surface. We are currently studying methods to smooth the surface prior to calculating the geometric descriptors. Employing a smoother surface should allow for using fewer shape descriptors, thus improving the performance of the algorithm.

CONCLUSIONS

We reviewed prior work with shape and introduced important categories of shape descriptors, surface versus bulk, and local versus global. We then introduced the novel QSD quadratic shape descriptor that effectively and efficiently represents the local surface shapes found in molecules. The descriptor is based upon a quadratic approximation to small regions of the molecular surface. We have shown how the QSD provides an invariant "surface centered" coordinate system that facilitates molecular alignment to superimpose two local shape descriptors. We presented our QSD algorithm for rapid *subshape similarity searching* that capitalizes on the beneficial properties of this new shape descriptor and showed that the algorithm is able to superimpose molecules

that are dissimilar in structure and in size more rapidly than any algorithm previously reported. We also showed that it can align shape similar cavities in proteins. The great speed of the QSD algorithm is attributed to its low order of worst case complexity, $O(nm^2)$. Finally, we showed that shape filtering using the difference in shape, $\Delta S = 0.3$, further reduced the complexity by about 70%. Other applications will be described in later papers in this series.

ACKNOWLEDGMENT

This work was supported by a Digital Equipment Corp. equipment grant, by National Institutes of Health AIDS Program Project Grant PO1 GM39599 shared with Agouron Pharmaceutical, and by UC Biotech Training Grant SC95-38. We also wish to thank Zhen Wang for his assistance in the active site alignment application in this work, supported by University of California BioSTAR 598-45.

REFERENCES AND NOTES

(1) Johnson, M. A.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; John Wiley & Sons, Inc.: New York, 1990.
(2) Dean, P. M., Ed. *Molecular Similarity in Drug Design*; Blackie Academic & Professional: New York, 1995.
(3) Martin, Y. C.; Bures, M. G.; Danaher, E. A. A Fast New Approach to Pharmacophore Mapping and Its Application to Dopaminergic and Benzodiazepine Agonists. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 83−102.
(4) Marshall, G. R. Binding-Site Modeling of Unknown Receptors. In *3D QSAR in Drug Design. Theory, Methods and Applications*; Kubinyi, H., Ed.; Escom: Leiden, The Netherlands, 1993; pp 80−116.
(5) Bures, M. G.; Danaher, E. A.; Martin, Y. C. J. D. New Molecular Modeling Tools Using Three-Dimensional Chemical Structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 218−223.
(6) Jones, G.; Willett, P.; Glen, R. C. A Genetic Algorithm for Flexible Molecular Overlay and Pharmacophore Elucidation. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 532−549.
(7) Gund, P.; Wipke, W. T.; Langridge, R. Computer Searching of a Molecular Structure File for Pharmacophoric Patterns. In *Computers in Chemical Research and Education*; Elsevier: Amsterdam, 1973; Vol. II, pp 5/33−38.
(8) Hahn, M. Three-Dimensional Shape-Based Searching of Conformationally Flexible Compounds. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 80−86.
(9) Sprague, P. W. Automated Chemical Hypothesis Generation and Database Searching with Catalyst. *Perspect. Drug Discovery Des.* **1995**, *3*, 1−20.
(10) Wild, D. J.; Willett, P. Similarity Searching in Files of Three-Dimensional Chemical Structures−Alignment of Molecular Electrostatic Potential Fields with a Genetic Algorithm. *J. Chem. Inf Comput. Sci.* **1996**, *36*, 159−167.
(11) Martin, Y. C. 3D Database Searching in Drug Design. *J. Med. Chem.* **1992**, *35*, 2145−2154.
(12) Artymiuk, P. J.; Bath, P. A.; Grindley, H. M.; Pepperrell, C. A.; Poirrette, A. R.; Rice, D. W.; Thorner, D. A.; Wild, D. J.; Willett, P. Similarity Searching in Databases of Three-Dimensional Molecules and Macromolecules. *J. Chem. Inf. Comput. Sci.* **1992**, *36*, 617−630.
(13) Burke, B. J.; Hopfinger, A. J. Advances in Molecular Shape Analysis. In *3D QSAR in Drug Design. Theory, Methods and Applications*; Kubinyi, H., Ed.; Escom: Leiden, The Netherlands, 1993; pp 276−306.
(14) Wipke, W. T. MACCS and REACCS. *Proceedings of the Society of Polymer Science*, Tokyo, Japan; 1983; pp 14−19.
(15) Wipke, W. T.; Nourse, J. G.; Moock, T. Generic Queries in the MACCS System. In *Computer Handling of Generic Chemical Structures*; Barnard, J. M., Ed.; Gower: Hampshire, England, 1984; pp 167−178.
(16) Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17−21.
(17) Randic, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1974**, *97*, 6609−6615.
(18) Randic, M. On Computation of Optimal Parameters for Multivariate Analysis of Structure Property Relationships. *J. Comput. Chem.* **1991**, *12*, 970−980.
(19) Rouvray, D. H. The Modeling of Chemical Phenomena Using Topological Indices. *J. Comput. Chem.* **1986**, *8*, 470−480.
(20) Basak, S.; s, S. B.; Grunwald, G. Application of Graph Theoretical Parameters in Quantifying Molecular Similarity and Structure Activity Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 270−276.
(21) Bemis, G. W.; Kuntz, I. D. A Fast and Efficient Method for 2D and 3D Molecular Shape Descriptors. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 607−628.
(22) Nilakantan, R.; Bauman, N.; Venkataraghavan, R. New Method for Rapid Characterization of Molecular Shapes: Applications in Drug Design. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 79−85.
(23) Bath, P. A.; Poirette, A. R.; Willet, P. Similarity Searching in Files of Three-Dimensional Chemical Structures: Comparison of Fragment-Based Measures of Shape Similarity. *J. Chem. Inf. Comput. Sci.* **1994**, *36*, 141−147.
(24) Pepperrell, C. A.; Willett, P. Techniques for the Calculation of 3-Dimensional Structural Similarity Using Interatomic Distances. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 455−474.
(25) Good, A. C.; Ewing, T. J. A.; Gschwend, D. A.; Kuntz, I. D. New Molecular Shape Descriptors: Application in Database Screening. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 1−12.
(26) Hermann, R. B.; Herron, D. K. OVID and SUPER: Two Overlap Programs for Drug Design. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 511−524.
(27) Barakat, M. T.; Dean, P. M. Molecular Structure Matching by Simulated Annealing. 1. A Comparison Between Different Cooling Schedules. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 295−316.
(28) Perkins, T. D. H.; Mills, J. E. J.; Dean, P. M. Molecular Surface-Volume and Property Matching to Superpose Flexible Dissimilar Molecules. *J. Comput.-Aided Mol. Design* **1995**, *9*, 479−490.
(29) Thorner, D. A.; Wild, D. J.; Willett, P.; Wright, P. M. Similarity Searching in Files of Three-Dimensional Chemical Structures: Flexible Field-Based Searching of Molecular Electrostatic Potentials. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 900−908.
(30) Chau, P. L.; Dean, P. M. Molecular Recognition: 3D Surface Structure Comparison by Gnomonic Projection. *J. Mol. Graphics* **1987**, *5*, 97−100.
(31) 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.
(32) Bladon, P. A Rapid Method for Comparing and Matching the Spherical Parameter Surfaces of Molecules and Other Irregular Objects. *J. Mol. Graphics* **1989**, *7*, 130−135.
(33) van Geerestein, V. J.; Perry, N. C.; Grootenhuis, P. D. J.; Haasnoot, C. A. G. 3D Database Searching of the Basis of Ligand Shape Using the SPERM Prototype Method. *Tetrahedron Comput. Methodol.* **1990**, *3* (6c), 595−613.
(34) Nissink, J. W. M.; Verdonk, M. L.; Kroon, J.; Mietzner, T.; Klebe, G. Superposition of Molecules: Electron Density Fitting by Application of Fourier Transforms. *J. Comput. Chem.* **1997**, *18* (5), 638−645.
(35) Blaney, F.; Finn, P.; Phippen, R.; Wyatt, M. Molecular Surface Comparison. *J. Mol. Graphics* **1993**, *11*, 98−105.
(36) Masek, B. B.; Merchant, A.; Matthews, J. B. Molecular Skins: A New Concept for Quantitative Shape Matching of a Protein with Its Small Molecule Mimics. *Proteins* **1993**, *17*, 193−202.
(37) Lawton, J.; Tudor, M.; Wipke, W. T. "The Basic Shape Topology of Protein Interfaces". In *Rational Drug Design*; Parrill, A., Reddy, M. R., Eds.; ACS Symposium Series 719; American Chemical Society: Washington, DC, 1999; pp 239−253.
(38) Lawton, J. N.; Wipke, W. T. QSD Quadratic Shape Descriptors. 5. A General Scoring Algorithm for Shape Comparison Applied to Evaluating Shape Complementarity as a Function of Molecular Size. Manuscript in preparation.
(39) Goldman, B. B.; Wipke, W. T. QSD Quadratic Shape Descriptors. 2. Molecular Docking Using Quadratic Shape Descriptors (QSDock). *Proteins* **2000**, *38*, 79−94.
(40) Goldman, B. B.; Wipke, W. T. QSD Quadratic Shape Descriptors. 3. Flexible Molecular Docking. Manuscript in preparation.
(41) Carson, M. Wavelets and Molecular Structure. *J. Comput.-Aided Mol. Design* **1998**, *10*, 273−283.
(42) Leicester, S. E.; Finney, J. L.; Bywater, R. P. Description of Molecular Surface Shape using Fourier Descriptors. *J. Mol. Graphics* **1988**, *6*, 104−108.
(43) Duncan, B. S.; Olson, A. J. Approximation and Characterization of Molecular Surfaces. *Biopolymers* **1993**, *53*, 219−229.
(44) Colloc'h, N.; Mornon, J.-P. A New Tool for the Qualitative and Quantitative Analysis of Protein Surfaces using B-spline and Density of Surface Neighborhood. *J. Mol. Graphics* **1990**, *8*, 133−140.
(45) Hosaka, M. *Modeling of Curves and Surface in CAD/CAM*; Springer-Verlag: New York, 1992; Chapter 6, pp 67−77.
(46) Connolly, M. L. Molecular Surface-Dot (MS, QCPE 429). *QCPE Bull.* **1981**, *1*, 18.

(47) Lin, S. L.; Fischer, D.; Nussinov, R.; Wolfson, H. J. Molecular Surface Representation by Sparse Critical Points. *Proteins* **1994**, *18*, 94−101.
(48) Connolly, M. L. Analytical Molecular Surface Calculation. *J. Appl. Crystallogr.* **1983**, *16*, 548−558.
(49) Zachmann, C. D.; Heiden, W.; Schlenkrich, M.; Brickmann, J. Topological Analysis of Complex Molecular Surfaces. *J. Comput. Chem.* **1992**, *13* (1), 76−84.
(50) Koenderink, J. J. *Solid Shape*; The MIT Press: Cambridge, MA, 1990; Chapter 6, pp 195−340.
(51) Wetherill, G. B.; Duncombe, P.; Kenward, M.; Kollerstrom, J.; Paul, S. R.; Vowden, B. J. *Linear Regression Analysis with Applications*; Chapman and Hall: New York, 1986; Chapter 1, pp 1−16.
(52) Boas, M. L. *Mathematical Methods in the Physical Sciences*; Wiley: New York, 1966; Chapter 3, p 106.
(53) Nussinov, R.; Wolfson, H. Efficient Detection of Three Dimensional Structural Motifs in Biological Macromolecules by Computer Vision Techniques. *Proc. Natl. Acad. Sci. U.S.A.* **1990**, *6*, 578−589.
(54) Norel, R.; Lin, S. L.; Wolfson, H. J.; Nussinov, R. Molecular Surface Complementarity at Protein-Protein Interfaces: The Critical Role Played by Surface Normals at Well Placed, Sparse, Points in Docking. *J. Mol. Biol.* **1995**, *252*, 263−273.
(55) Rarey, M.; Wefing, S.; Lengauer, T. Placement of Medium-Sized Molecular Fragments into Active Sites of Proteins. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 41−54.
(56) Lin, S. L.; Nussinov, R. Molecular Recognition via Face Center Representation of a Molecular Surface. *J. Mol. Graphics* **1996**, *14*, 78−90.
(57) Korolkovas, A. *Essentials of Molecular Pharmacology; Background for Drug Design*; Wiley: New York, 1970.
(58) Connolly, M. L. Solvent-Accessible Surfaces of Proteins and Nucleic Acids. *Science* **1983**, *221*, 709−713.
(59) Stouch, T. R.; Jurs, P. C. A Simple Method for the Representation, Quantification, and Comparison of the Volumes and Shapes of Chemical Compounds. *J. Chem. Inf Comput. Sci.* **1986**, *26*, 4−12.
(60) Hopfinger, A. J. A QSAR Investigation of DHFR Inhibition by Baker Triazines Based upon Molecular Shape Analysis. *J. Am. Chem. Soc.* **1980**, *102*, 7196−7206.
(61) Lawton, J. N.; Goldman, B. B.; Wipke, W. T. QSD Quadratic Shape Descriptors. 6. An Information Theoretic Approach to Assessing Ligand-Receptor Shape-Complementarity. Manuscript in preparation.
(62) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasmui, M. The Protein Data Bank: A Computer-Based Archival File for Macromolecular Structures. *J. Mol. Biol.* **1977**, *112*, 535−542.
(63) Bohm, H.-J.; Klebe, G. What Can we Learn from Molecular Recognition in Protein-Ligand Complexes for the Design of New Drugs. *Angew. Chem., Int. Ed. Engl.* **1996**, *35*, 2588−2614.
(64) Loll, P. J.; Picot, D.; Ekabo, O.; Garavito, R. M. Synthesis and Use of Iodinated Nonsteroidal Antiinflammatory Drug Analogs as Crystallographic Probes of the Prostaglandin H-2 Synthase Cyclooxygenase Active Site. *Biochemistry* **1996**, *35*, 7330−7340.
(65) Kurumbail, R. G.; Stevens, A. M.; Gierse, J. K.; McDonald, J. J.; Stegeman, R. A.; Pak, J. Y.; Gildehaus, D.; Miyashiro, J. M.; Penning, T. D.; Siebert, K.; Isakson, P. C.; Stallings, W. C. Structural Basis for Selective Inhibition of Cyclooxygenase-2 by Anti-Inflammatory Agents. *Nature* **1996**, *384*, 644−648.
(66) Jorgensen, W. Rusting of the Lock and Key Model for Protein-Ligand Binding. *Science* **1991**, *254* (5034), 954−955.

CI980213W