A Novel Shape-Feature Based Approach to Virtual Library Screening

Santosh Putta,* Christian Lemmen,[‡] Paul Beroza,[§] and Jonathan Greene[#]

Deltagen Research Labs, 740 Bay Road, Redwood City, California 94063-2469

Received January 8, 2002

The shape of and the chemical features of a ligand are both critical for biological activity. This paper presents a strategy that uses these descriptors to build a computational model for virtual screening of bioactive compounds. Molecules are represented in a binary shape-feature descriptor space as bit-strings, and their relative activities are used to identify the subset of the bit-string that is most relevant to bioactivity. This subset is used to score virtual libraries. We describe the computational details of the method and present an example validation experiment on thrombin inhibitors.

1. INTRODUCTION

Our understanding of the interactions responsible for ligand−receptor binding has been greatly aided by X-ray crystallography, which provides the three-dimensional coordinates for many molecules of pharmacological interest. Surveys of crystal structures for ligands,[1,2] protein−protein complexes,[3,4] and protein−ligand complexes[5,6] have shown that two factors are of primary importance in determining binding affinity: (1) shape complementarity and (2) complementary physical-chemical features; chemical groups on the ligand and receptor are arranged in such a way that the energy of the complex is lowered by chemical interactions.

Computational methods to predict the binding affinities of small molecules to proteins whose three-dimensional structure is known make direct use of these two factors: conformations of a small molecule are explored for those that complement the shape of the receptor site. Those molecules that fit well are scored on the basis of the physicochemical complementarity between the docked conformer and the receptor.[7−11] However, such methods are possible only when the protein structure is available in atomic detail. When that information is unavailable, an indirect approach that focuses on ligand-based analysis is required. In such cases, the general strategy is to compare ligands to each other with the hope that some common shape and chemical features can be identified to explain the activities of the molecules and, further, that these commonalities are generally applicable to molecules for which activity data is unknown.

We present such a ligand-based approach that relies on comparisons of ligand shapes and the locations of chemical features within those shapes. The shapes a ligand can attain are enumerated, and the locations of the ligand's chemical features within these shapes are identified. This information is encoded into a binary string: the shape-feature signature. The signature is essentially a shape based descriptor, though it is significantly different from other shape based descriptors in the literature.[12−14] Shape-feature signatures of molecules with known biological activity are then evaluated to construct models for biomolecular activity.

This paper explains the computational methods used to define and compare molecular shapes and the construction and analysis of shape-feature signatures. We present some example applications and results; more detailed applications of the method have been discussed in a companion paper.[15] This publication is structured as follows: section 2 presents an overview of the methodology and section 3 is a thorough description of the technical details. Section 4 presents a way of assessing the performance of the shape alignment aspects of the method. Section 5 describes results on a thrombin data set. The overview section provides sufficient information to proceed to the results section (section 5), if the reader wishes to avoid the computational details.

2. METHODOLOGY: OVERVIEW

The goal of this method is to obtain a model for activity that consists of a set of shapes and chemical features in these shapes. The data set given as input to the method consists of compounds with measured affinities against a biologically relevant target. We assume that an ensemble of three-dimensional conformations for each compound of interest is available. The method proceeds in four stages (Figure 1): defining a descriptor space, mapping molecules into the descriptor space, building a model for bioactivity, and selecting compounds from a virtual library based on the model.

Because the shape of the bioactive conformation for each active molecule is unknown, the method starts by considering a large set of shapes that are potentially relevant for activity. These shapes along with the possible locations for chemical features within them constitute what we call the *descriptor space*. Defining a descriptor space starts by choosing a *target set* of molecules from the data set. Typically the target set consists of the active molecules, a portion of whose conformers are likely to produce shapes that are good candidates for a model for bioactivity. The target set is then

* Corresponding author phone: (650)569-5305; fax: (650)569-5502; e-mail: sputta@combichem.com.

[‡] Current address: BiosolveIT GmbH, An der Ziegelei 75, 53757 Sankt Augustin, Germany.

[§] Current address: Telik Inc., 750 Gateway Blvd, South San Francisco, CA 94080.

[#] Current address: 1481 Pitman Ave, Palo Alto, CA 94301.

10.1021/ci0255026 CCC: $22.00
Published on Web 09/06/2002

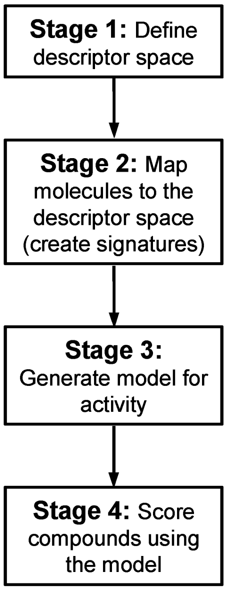
Stage 1: Define descriptor space
Stage 2: Map molecules to the descriptor space (create signatures)
Stage 3: Generate model for activity
Stage 4: Score compounds using the model

Figure 1. The four stages of shape-feature model generation and virtual library screening.

used to create a *shape catalog*. The shape catalog is a diverse collection of shapes that the conformations of the molecules in the *target set* can attain. These shapes are generated by processing the compounds in the target set in a random order. Each compound's conformers are processed sequentially. Since the shape catalog is initially empty, the shape of the first conformation of the first compound is by default added to the catalog. From then on the shape of a conformer is compared to the population of shapes in the shape catalog. If the conformer's shape is sufficiently similar to any of those shapes that are already present in the catalog, it is considered to be represented in the catalog. However, if it is sufficiently different from each shape in the catalog, the conformer's shape is added to the catalog. After consideration of the shapes for all conformations of all molecules in the target set, the shape catalog is complete.

The additional component of the descriptor space is the location of chemical features within each shape. By chemical features, we mean the set of molecular fragments typically associated with ligand−receptor binding (i.e., hydrogen bond donors and acceptors, groups bearing a net electrostatic charge, hydrophobic groups, and aromatic rings). The descriptor space now consists of a set of shapes and the locations of the chemical features within them. The shape-feature descriptor itself is represented by a binary string (i.e., a vector of ones and zeros). Each bit in the signature encodes the following: shape, feature type, and grid location of the feature.

Once the descriptor space has been defined, each molecule needs to be represented in it, i.e., a shape-feature signature is generated for each molecule. The conformers for each compound in the data set are compared to each of the shapes in the shape catalog. If a conformer matches a shape in the catalog, the location of each of its chemical features (hydrogen bonding groups, hydrophobic groups, charges, and aromatic ring centers) is translated into the bits associated with that shape, feature type, and feature location. The corresponding bits (descriptors) in the binary string are turned on. All other bits are pre-set and remain zero. This process is repeated for all the conformers of a compound. Note here that the bits are set in the same signature for all the conformers of a compound. In other words, the signature is the cumulative description of a compound over its conformers. Each feature and shape combination (i.e., each bit in the signature) is called a *hypothesis* for activity. We use the terms bit (signature bit) and hypothesis interchangeably. Once all the molecules have been processed, we have an array of binary signatures that we process to build models for biological activity.

The third phase of the method takes the binary encoding of all of the input compounds and identifies those bits that are relevant for activity. A subset of these bits is selected as an *ensemble hypothesis*. This ensemble defines the model for activity derived from the input data set. It usually comprises hypotheses from a limited number of shapes typically 10−30 and a number of feature locations relative to these shapes.

The final stage of the method consists of using the model to score a set of compounds by the number of ensemble bits they match. This requires that shape feature signatures be generated for compounds in the virtual library. In the scoring phase, however, only shapes represented in the ensemble need to be considered for matching the conformers of a compound. Once the compounds have been selected by the computational model, they are submitted for biological assay.

3. METHODOLOGY: DETAILED

This section explains each of the stages mentioned in the previous section in more detail. The method takes as input a set of precomputed conformations for each molecule in the data set. We use our in-house tool CONAN,[16,17] which was designed to generate a representative set of low-energy conformations for a molecule. Typically 30−200 low-energy conformations are used per compound.

3.1. Shape Catalog Generation. This subsection is divided into four parts. The first three describe the shape catalog generation procedure, shown in Figure 2, and the last section describes a procedure to optimize the speed of the shape catalog generation.

3.1.1. Canonical Posing of Conformers. The first step in obtaining a shape from a conformer is to pose it in a standard way, called the *canonical orientation*. Our definition of the canonical orientation depends on the presence or absence of a *key-feature*. A key-feature is a chemical group common to all active molecules that can serve as a starting point for alignment. For example, if all the active molecules contain a positively charged moiety, we can translate each molecule so that the positive charge coincides with the origin of the coordinate axes.

In the absence of a key-feature, the centroid of a conformer is used as the point of alignment. The centroid is calculated from the equally weighted coordinates of its heavy (i.e., nonhydrogen) atoms. If $\vec{p}_i$ is the position vector of the *i*th heavy atom, $\vec{p}_c$ the position vector of the centroid is calculated as

$$\vec{p}_c = \frac{1}{n}\sum_{i=1}^{n}\vec{p}_i \qquad (1)$$

where *n* is the total number of heavy atoms in the compound.

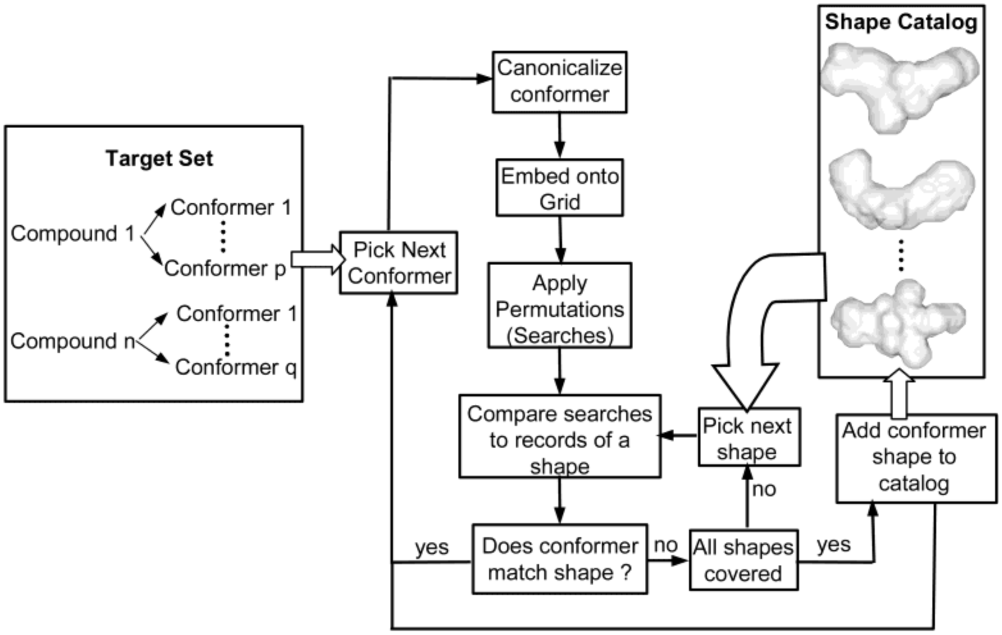

Figure 2. Shape catalog is typically generated from a target set of molecules (typically actives). Each conformer shape that is not already represented in the catalog is added to the catalog until all conformers in the target set have been evaluated.

The principal axes of the conformer about the centroid are determined by computing the eigenvectors of the following covariance matrix

$$C = \begin{bmatrix} \sum_{i=1}^{n} x_i x_i & \sum_{i=1}^{n} x_i y_i & \sum_{i=1}^{n} x_i z_i \\ \sum_{i=1}^{n} y_i x_i & \sum_{i=1}^{n} y_i y_i & \sum_{i=1}^{n} y_i z_i \\ \sum_{i=1}^{n} z_i x_i & \sum_{i=1}^{n} z_i y_i & \sum_{i=1}^{n} z_i z_i \end{bmatrix} \qquad (2)$$

where x_i, y_i, and z_i are the coordinates of the ith heavy atom. The corresponding eigenvalues when divided by the number of heavy atoms are an approximation of the second-order moments of the conformer shape. We refer to these moments as the *even moments* denoted by $M[X^2]$,$M[Y^2]$,$M[Z^2]$. Note that we use upper case for the coordinates in order to indicate that they correspond to values after the conformation has been centered and its moments aligned with the coordinate axes. The conformer is transformed such that the centroid coincides with the origin and the eigenvectors align with the x, y, and z axes. At this stage, however, the choice of the positive and negative sense of the axes is arbitrary. For example, the even moments do not change if we rotate about the third principal axis such that the first principal axis is replaced by its negative. Such *flipping* is possible for the second principal axis as well. However, the orientation of the third principal axis is completely determined (to maintain a right-handed coordinate system), once the first two axes are chosen. A consistent choice of the axis flips can be obtained using the third-order moments, calculated using the following formulas:

$$M[X(Y^2 + Z^2)] = \sum_{i=1}^{n} X_i(Y_i^2 + Z_i^2) \qquad (3)$$

$$M[Y(X^2 + Z^2)] = \sum_{i=1}^{n} Y_i(X_i^2 + Z_i^2) \qquad (4)$$

$$M[Z(X^2 + Y^2)] = \sum_{i=1}^{n} Z_i(X_i^2 + Y_i^2) \qquad (5)$$

We refer to these moments as the *odd moments*. The odd moments carry information regarding the symmetry of the conformer shape about the x, y, and z axes. For example, in Figure 3 the first shape is symmetric about the yz-plane and the odd moment $M[X(Y^2 + Z^2)]$, is close to zero. However, the odd moment $M[Y(X^2 + Z^2)]$ is large and positive, reflecting the asymmetry about the xz-plane. Also in Figure 3 the positive value of $M[Y(X^2 + Z^2)]$ for the first shape indicates the upward concave nature of the shape. The moments for the second shape in the Figure 3 give similar information regarding the symmetry about the x-axis and the asymmetry about the yz-plane. The positive side of the first principal axis is chosen such that $M[X(Y^2 + Z^2)]$ is positive. A similar criteria is applied to the second principal axis to make the second odd moment $M[Y(X^2 + Z^2)]$ positive.

In the presence of a key-feature, the location of the key-feature is used as the origin, and the canonical orientation is obtained as follows. The first axis is chosen such that the centroid of the conformer falls on the x-axis. Then, for all of the heavy atoms, the perpendicular distance d from the atom to the x-axis is calculated. The conformer is rotated about the x-axis such that the heavy atom with the largest d value falls in the xy plane. After this transformation, the even and odd moments given by eqs 3−6 are calculated. Just as in the previous case, the x-axis and y-axis may be flipped by 180° to make the odd moments $M[X(Y^2 + Z^2)]$ and $M[Y(X^2 + Z^2)]$ positive.

The alignment procedures described above are referred to as *centroid* and *key-feature alignment*, respectively.

3.1.2. Encoding Grid Occupancy. Once a conformer orientation has been canonicalized, it is superimposed onto

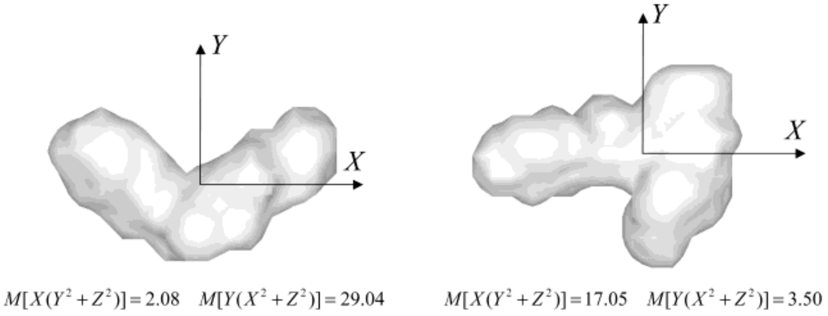

Figure 3. The third-order moments $M[X(Y^2 + Z^2)]$, $M[Y(X^2 + Z^2)]$, $M[Z(X^2 + Y^2)]$ can be used both to get an approximate idea of the shape and to orient the shape consistently. In the first picture the $M[X(Y^2 + Z^2)]$ is close to zero indicative of the symmetry of the shape about the *yz*-plane. However, $M[Y(X^2 + Z^2)]$ is large, indicating the asymmetry about the *xz*-plane. In the second picture the shape is more symmetric about the *xz*-plane than the *yz*-plane and as a result $M[X(Y^2 + Z^2)]$ is large, while $M[Y(X^2 + Z^2)]$ is close to zero.

Table 1. Occupancy at Each Grid Point Is Assigned Based on Its Distance to the Closest Heavy Atom[a]

range of d	grid occupancy value
$d \le r_w$	3
$r_w < d \le \left(r_w + \frac{r_b}{s_g}\right)$	2
$\left(r_w + \frac{r_b}{s_g}\right) < d \le \left(\frac{2.r_b}{s_g}\right)$	1
$d > \left(\frac{2.r_b}{s_g}\right)$	0

[a] d is the distance of the grid point to the center of closest heavy atom. r_w is the van der Waal radius of the heavy atom. r_b is a user specified parameter. s_g is the grid spacing.

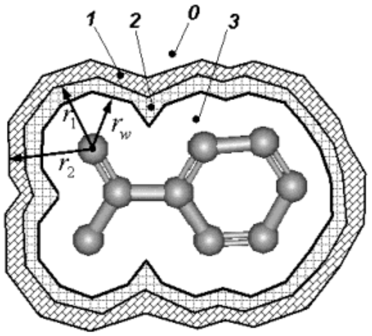

Figure 4. A 2D illustration of grid encoding. All grid points in the white region at the center get a value of 3. The value decreases gradually, as shown by the shaded regions, as one moves beyond the van der Waal surface of the molecule. In the figure r_w is the van der Waal. $r_1 = (r_w + (r_b/s_g))$ and $r_2 = (r_w + (2.r_b/s_g))$, where r_b is a user specified parameter. s_g is the grid spacing.

a 3D cubic lattice (grid) to facilitate encoding its shape. The grid dimensions are chosen to accommodate all compounds in their canonical orientation. The edge length of each cube in the lattice is specified by the user and is referred to as the *grid spacing*, which determines the accuracy of the shape encoding. Note that the grid is identical for every shape. The grid occupancy of a conformer is encoded using a bit vector whose length is proportional to the number of grid points. Two bits are assigned to store the occupancy of an individual grid point. This allows for four distinct values based on whether the grid point is interior, exterior or close to the surface of the conformer shape. Table 1 shows the encoding scheme based on the distance of a grid point to the closest heavy atom and the van der Waals radius of the atom. Figure 4 shows a 2D illustration of the encoding for a typical molecule. The idea of using four levels for encoding a grid point is borrowed from the *anti-aliasing* technique used in computer graphics.[18] The accuracy achieved this way is comparable to a two value encoding on a grid with half the grid spacing. However, in the latter case 8 times as many grid points and 4 times as many bits are required.

We use the Tanimoto distance, d_t, as a measure of distance or dissimilarity between shapes

$$d_t = 100\cdot\left(1 - \frac{v_1 \cap v_2}{v_1 \cup v_2}\right) \tag{6}$$

where $v_1 \cap v_2$ is the overlap (common) volume occupied by the two shapes, and $v_1 \cup v_2$ is the union volume occupied by the two shapes. The grid occupancy bit vectors are used to compute both these quantities rapidly. However, a direct bit-wise union and intersection operation will not yield the correct answers, because of the 2 bits per grid point encoding used for the shapes. As a result we use a customized procedure that uses a lookup table to accurately compute the union and intersection volumes between the grid occupancy vectors.

3.1.3. Shape Accumulation. As each molecule's conformation is transformed into an array of grid values, it is compared to those shapes already encountered and retained if it is sufficiently unique. The accumulated set of grid-encoded shapes thus retained is referred to as the shape catalog. Usually, the shape catalog is generated from the conformers of the active molecules, processed in random order. When a shape is added to the catalog it is stored in several orientations. The shape is first stored as an array of grid values corresponding to the conformer's canonical orientation. Comparing shapes of two conformations in their canonical orientation provided an optimal overlap in a significant number of cases (~25%). However, to increase the probability of optimal shape overlap, two types of rotations are applied to the conformer and the resulting pose is also stored as a separate array of grid values. Thus, when a new conformer is encountered it is compared to several poses of each shape in the catalog. The two different types of rotations applied to the conformer to perturb it from its canonical orientation are called *jiggle*s and *large rotations*. Jiggles are small rotations (typically 10 degrees) about a predefined set of uniformly spaced axes[19] (typically 10−20) in three-dimensional space. In cases in which two or more of the principal moments of the conformer are degenerate, a set of *large rotations* are applied in addition to the jiggles. If the moments $M[X^2]$ and $M[Y^2]$ satisfy

$$\left|\frac{M[X^2] - M[Y^2]}{M[X^2] + M[Y^2]}\right| < f \qquad (7)$$

where f is a user defined threshold (typically 0.1), large rotations are applied around the z-axis of the conformer. Equivalent criteria are applied to rotations about the x and y axes. *Large rotations* vary between 0 and 90° in user defined steps of 10°−30°. These rotations are mapped to the remaining three quadrants (90°−360°) by a third set of rotations called *permutations* (described below), which are applied to conformers before they are compared to shapes in the catalog. Typically using a value of 0.1 for f in inequality 7 above, the large rotations are necessary only for a small portion (<5%) of the shapes in the catalog.

Upon adding a conformer shape to the shape catalog, the combined rotations of a jiggle and a large rotation are applied to the conformer. The resulting orientation of the compound is superimposed onto the grid. The grid occupancy, defined this way, is stored in a bit vector, referred to as a *record*. For each shape in the catalog, all jiggle and large rotation combinations are stored as separate records. Thus, each shape is stored as a collection of grid occupancies that correspond to the grid representations of the shape after it has been modulated by the jiggles and large rotations. In addition to the set of records, the principal moments $M[X^2]$, $M[Y^2]$, $M[Z^2]$, $M[X(Y^2 + Z^2)]$, $M[Y(X^2 + Z^2)]$, $M[Z(X^2 + Y^2)]$ that are associated with the conformer that generated that shape are also stored for each shape in the catalog. Because the number of large rotations depends on the degeneracy of the principal moments of the conformer, shapes in the catalog may have a different number of records (i.e., grid representations of the shape).

Accumulation of shapes in the catalog occurs as follows. The shape of the first target conformer of the first molecule is always added to the shape catalog, since the catalog is empty at this point. After the first shape has been added to the shape catalog, the procedure followed to process the remaining target conformers is slightly different. A new target conformer is compared to all the shapes in the shape catalog. If it is sufficiently different from all of them, the shape of the new conformer is added to the catalog. The comparison process starts by applying a set of rotations called *permutations* to the conformer in its canonical orientation. Permutations are rotations that interchange the principal axes of the conformer. As mentioned earlier the canonical orientation aligns the first principal axis to the x-axis, the second to the y-axis and the third to the z-axis. A permutation assigns a different alignment of the first, second and third principal axes to the x, y, and z axes. There are 24 different permutations. Since a permutation changes only the axes, every permutation results in a unique one-to-one correspondence, which can be precomputed, of grid locations before and after the permutation. Hence the occupancy vector after the permutation can be computed simply by permuting the values in the original bit vector directly. These permuted bit vectors are called *searches*. Bit shuffling is obviously much faster than rotating a conformer and then computing the grid occupancy for the resulting orientation. Consequently the permutations are computed on the fly for each new conformer when required, rather than saving them with the shapes in the catalog. This procedure can save up to a factor of 24 in terms of the memory consumed by the shape catalog. Note that for an arbitrary orientation, like a jiggle or a large rotation, the one-to-one correspondence cannot be precomputed, hence, this shortcut does not apply. A similar one-to-one correspondence can be used to efficiently obtain the moments $M_c[X^2]$, $M_c[Y^2]$, $M_c[Z^2]$, $M_c[X(Y^2 + Z^2)]$, $M_c[Y(X^2 + Z^2)]$, $M_c[Z(X^2 + Y^2)]$, for a search from the moments $M[X^2]$, $M[Y^2]$, $M[Z^2]$, $M[X(Y^2 + Z^2)]$, $M[Y(X^2 + Z^2)]$, $M[Z(X^2 + Y^2)]$ of the conformer.

As in the case of large rotations, often not all of the permutations are necessary. Which ones to apply is determined by the moments $M[X^2]$, $M[Y^2]$, $M[Z^2]$. If a permutation interchanges the i-axis ($I = x$, y, or z) in the canonical orientation with the j-axis ($J = x$, y, or z), the corresponding search is computed only if the following condition holds

$$\left|\frac{M[I^2] - M[J^2]}{M[I^2] + M[J^2]}\right| < f \qquad (8)$$

where f is an user defined threshold, usually set at 0.1. The next step in the comparison process involves computing the Tanimoto distance (eq 6) between all searches of the new conformer and all the records of all the shapes in the shape catalog. Generally, it is not necessary to compute the Tanimoto distance between all the search-record pairs. The moments of the searches and the moments of the shapes in the catalog can be used to avoid unnecessary comparisons. This procedure, referred to as *moment filtering,* is detailed in the next subsection. If the Tanimoto distances between a record and a search is less than a user-defined threshold ($D_{\max}$), the conformer is said to *match* a shape in the catalog. In this case, the conformer shape is considered represented and discarded, and we move onto the next conformer. However, if no match is found, the conformer shape is added to the shape catalog. As described earlier, upon adding a shape to the shape catalog, all the necessary records (based on jiggles and large rotations) plus the moments are computed and stored. This whole procedure is repeated for the all target conformers.

3.1.4. Moment Filtering. As mentioned in the previous subsection, the moments can be used to avoid some unnecessary comparisons between searches and records, to save computing time. The moments can be thought of as rough descriptors of a shape. For example the even moments $M[X^2]$, $M[Y^2]$, $M[Z^2]$ are an approximation of the dimension of the shape along the x, y, and z axes. As a result a search is unlikely to match a record if their even moments are significantly different. As discussed earlier, odd moments $M[X(Y^2 + Z^2)]$, $M[Y(X^2 + Z^2)]$, $M[Z(X^2 + Y^2)]$ carry some information about the curvature/asymmetry of the conformer. Let $M_c[X^2]$, $M_c[Y^2]$, $M_c[Z^2]$, $M_c[X(Y^2 + Z^2)]$, $M_c[Y(X^2 + Z^2)]$, $M_c[Z(X^2 + Y^2)]$ be the even and odd moments of a search C_s of a conformer C, that is being compared to a record S_s of a shape S with moments $M_s[X^2]$, $M_s[Y^2]$, $M_s[Z^2]$, $M_s[X(Y^2 + Z^2)]$, $M_s[Y(X^2 + Z^2)]$, $M_s[Z(X^2 + Y^2)]$ in the shape

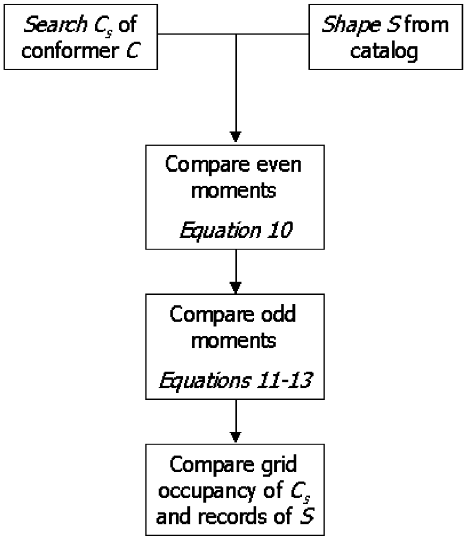

Figure 5. Computationally expensive comparisons between grid occupancies of the shape can be eliminated by using the second and third order moments of the shapes.

catalog. The even moments are used to check the following condition.

$$\left(\frac{M_c[X^2] - M_s[X^2]}{K_x}\right)^2 + \left(\frac{M_c[Y^2] - M_s[Y^2]}{K_y}\right)^2 + \left(\frac{M_c[Z^2] - M_s[Z^2]}{K_z}\right)^2 < 1.0 \quad (9)$$

The constants K_x,K_y,K_z, are user defined. The above inequality constrains the differences in the even moments to within an ellipsoid of axes lengths K_x,K_y,K_z. The ellipsoidal form and the values for K_x,K_y,K_z have been obtained empirically. If inequality 10 is violated we conclude that C_s and *S* do not match. Otherwise the following inequalities are evaluated

$$|M_c[X(Y^2 + Z^2)] - M_s[X(Y^2 + Z^2)]| < K_{x(y^2+z^2)} \quad (10)$$

$$|M_c[Y(X^2 + Z^2)] - M_s[Y(X^2 + Z^2)]| < K_{y(x^2+z^2)} \quad (11)$$

$$|M_c[Z(X^2 + Y^2)] - M_s[Z(X^2 + Y^2)]| < K_{z(x^2+y^2)} \quad (12)$$

where $K_{x(y^2+z^2)}$, $K_{y(x^2+z^2)}$, $K_{z(x^2+y^2)}$ are user defined constants. Only if all of the above conditions hold, is the pair C_s and *S* considered for the comparison of the bit vectors as described above. Figure 5 provides a flowchart of the moment filtering procedure.

3.2. Signature Generation. Signature generation, in which each molecule is mapped into a bit vector in the shape-feature descriptor space, is a two step procedure. The first step involves a shape comparison of the molecule's conformers with every shape in the shape catalog. In the second step, the signatures of the compounds whose conformers match any of the shapes in the catalog are updated by setting the bits in the signature that correspond to the matched shapes and feature locations within them.

3.2.1. Shape Comparison. The procedure for comparing a conformer to the shapes in the shape catalog is the same as the one used during the shape catalog generation (see Figures 2 and 6). First, a query conformer is transformed to its canonical orientation and then searches are computed depending on its moments. Note that there is a subtle distinction between the shape catalog generation phase and the signature generation phase. During shape catalog generation if no match is found for a conformer with any shape in the catalog, the conformer shape is added to the shape catalog. However, during signature generation, the shape catalog is left unchanged. If no single match is found for a conformer, we ignore this conformer and proceed to the next one. Only if a conformer matches a shape in the catalog is the signature of the corresponding compound updated.

Recall that multiple searches may be computed for each conformer and each shape in the catalog may have several records. As a result a conformer may match the same shape several times (in multiple orientations). In such cases the signature for the compound is updated each time.

3.2.2. Creating Signatures. Initially, a molecule's shape-feature signature consists entirely of zeros. When a match is found between one of the molecule's conformer's *search* bit vectors and a catalog shape's *record* bit vectors, the molecule's signature is updated by setting the appropriate bits in the molecule's signature. These bits are determined by where the molecule puts chemical features in the shape.

Chemical features are chemical substructures that are commonly associated with noncovalent binding: hydrogen bond donors and acceptors, positively and negatively charged groups, hydrophobic groups, and aromatic rings. These six chemical feature types are identified in each molecule through a set of queries in a SMARTS-like language.[20] Their definitions, including topological definitions for hydrophobicity, are assigned using rules similar to those previously reported.[21] Chemical features are represented in the descriptor space by superimposing a three-dimensional grid on each shape. By matching these queries to a compound we identify which features the molecule possesses and where the molecule puts them in three-dimensional space. For each conformer, the feature locations thus found are mapped onto a grid that is aligned with the shape grid and typically is of identical dimension and spacing.

The dimension, *N*, of the shape-feature binary signature is given by

$$N = N_s{\cdot}N_f{\cdot}N_g \quad (13)$$

where N_s is the number of shapes in the shape catalog. N_f is the number of chemical feature types, and N_g is the dimension of the feature grid. Each bit in the shape-feature signature encodes a feature of specific type, located at a specific grid point, on a specific shape. The location (address) *I* of a bit in the string that corresponds to the feature with index I_f at the I_gth grid point on the I_sth shape is given by the following equation:

$$I = (N_s{\cdot}N_f){\cdot}I_s + (N_g){\cdot}I_f + I_g \quad (14)$$

If a conformer matches a shape, the position of its features on the feature grid are calculated and the corresponding bits in the signature are turned on. Recall that a match is found between a particular search-record pair. Associated with these are a specific jiggle and large rotation and a specific permutation, respectively. Thus a defined transformation has

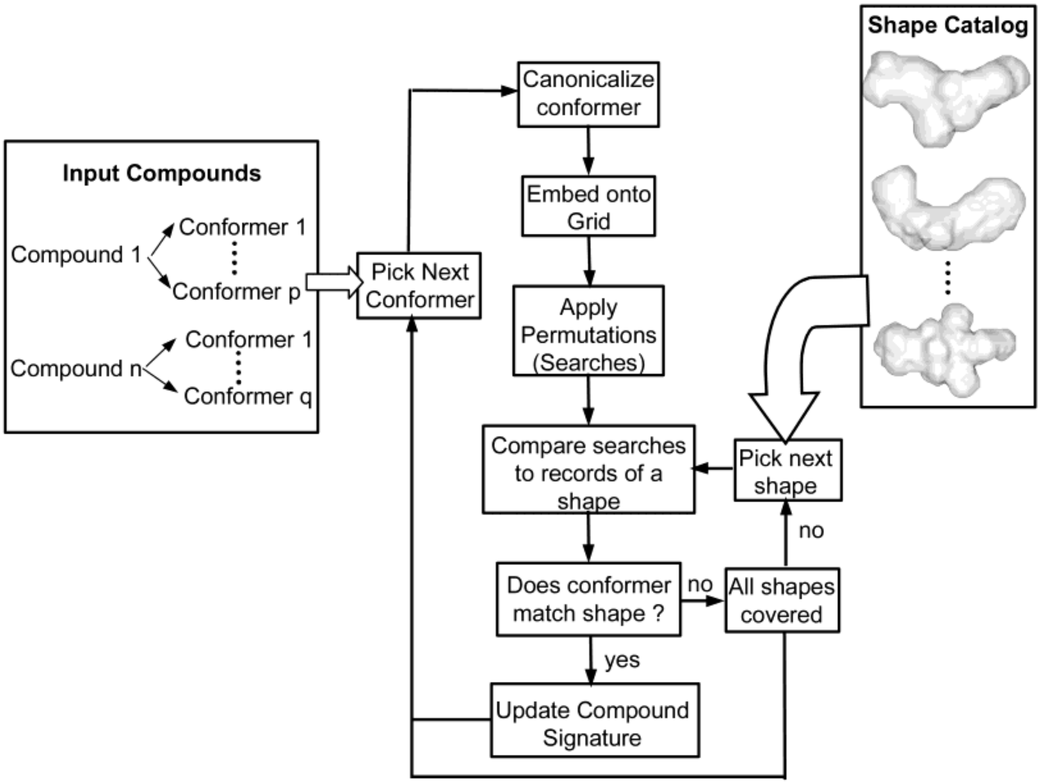

Figure 6. The procedure followed for generating signatures is very similar to the one followed for generating the shape catalog (see Figure 2). However, during signature generation the shape catalog remains unchanged, and the signature for the compound is updated every time one of its conformers matches a shapes in the catalog.

to be applied to obtain the match. The exact same transformation is applied to the feature locations to position them on the feature grid. Each conformation of a molecule in turn is used to set bits in the molecule's signature. Hence, a molecule's signature represents the union of the shape-feature combinations of all of its conformers.

Note that shape-feature signatures are usually quite sparse. Molecules typically do not match a large number of shapes and have only a limited number of features, resulting in very few bits being set. Our signatures typically contain around one million bits, out of which less than 25 000 are typically turned on for a drug-like molecule.

3.3. Processing Signatures. The next step is to build a model for bioactivity by selecting those hypotheses that best discriminate active from inactive compounds. The search space here typically contains around 1 000 000 bits, most of which are essentially *uninformative*. We use a program called SIGNAL[22] that performs both the selection of informative hypotheses and facilitation of a cross-validation protocol. SIGNAL uses a mutual information function to select the most informative hypotheses

$$IC = -h\left(\frac{N_a}{N}\right) - h\left(\frac{N_i}{N}\right) + h\left(\frac{N_{ap}}{N_p}\right) + h\left(\frac{N_{ip}}{N_p}\right) + h\left(\frac{N_{an}}{N_n}\right) + h\left(\frac{N_{in}}{N_n}\right) \tag{15}$$

where $h(x) = x \log_2(x)$, the subscripts a and i refer to active and inactive compounds in the data set, respectively, and the subscripts p and n refer to the hypothesis-matching and not hypothesis-matching compounds, respectively. N refers to the total number of compounds in the data set. This function favors hypotheses that are present in many active compounds and penalizes hypotheses that are present in many inactive compounds.

A maximally informative hypothesis is one that is present in all actives ($N_{ap} = N_a$) and none of the inactives ($N_{ip} = 0$). In this case

$$\frac{N_{ip}}{N_p} = \frac{N_{an}}{N_n} = 0 \quad \text{and} \quad \frac{N_{ap}}{N_p} = \frac{N_{in}}{N_n} = 1 \tag{16}$$

Here we follow the convention that $h(0) = 0$. Since $h(1) = 0$, it follows that $I = -h(N_a/N) - h(N_i/N)$, which is equivalent to the entropy in the data set, and it is also the maximum information that a hypothesis can possibly have. At the other extreme, a hypothesis that is present in all actives and all inactives gives $N_p = N$, $N_{ap} = N_a$, $N_{ip} = N_i$, and $N_{an} = N_{in} = 0$. Thus, the first and the third as well as the second and the fourth term of eq 17 cancel each other out, and $h(N_{an}/N_n) = h(N_{in}/N_n) = 0$, by convention. Thus $I = 0$, which means the hypothesis provides no information about activity. The top-ranking hypotheses (typically, up to 100) define an *ensemble*, which is our model for activity. The reader is reminded here that each of the hypotheses (or bit) in the ensemble contains a unique encoding of the shape, chemical feature and location of the feature on the grid. It is possible that each of the hypotheses corresponds to a different shape in the catalog (e.g. 100 hypotheses from 100 shapes). However, it has been observed empirically that this is rarely the case. Most often the ensemble hypotheses belong only to a small portion of the shapes. For example 100 hypotheses in the ensemble typically belong to about 20 shapes. The hypotheses that belong to the same shape are different only by virtue of the feature type and/or the location of the feature on the grid.

3.4. Virtual Library Screening. The final stage in our shape-feature based approach is to use the ensemble model to screen virtual libraries for biologically promising compounds. To score compounds in a virtual library using this approach, it is first necessary to construct shape-feature signatures for the molecules in the library. The procedure is identical the signature generation for the molecules of known activity, with one exception: that shape catalog can be reduced to only those shapes that are represented by the bits in the ensemble hypothesis; the other shapes were not strongly associated with activity and may be discarded. Once the signatures for the molecules in the virtual library have been constructed, the compounds are scored by the number of ensemble hypotheses that the compound matches. In other words, the score reflects how many of the most informative features in corresponding locations are met by the compound. A cross-validation protocol is used to determine a threshold for the score that is most likely to distinguish, optimally, active from inactive compounds.

3.4.1. Multiple Hypotheses per Shape. As mentioned earlier the ensemble hypotheses belong only to a small subset of shapes in the catalog. These shapes are referred to as the *ensemble shapes*. Generally speaking, the ensemble shapes are considered to be potentially important for activity, as suggested by the data. As mentioned in the previous section, a compound is predicted by our model to be active if more than a threshold number of the ensemble hypotheses are turned on in the compound signature. Note that a compound with *x* ensemble bits turned on may match only one ensemble shape and *x* features on that shape. At the other extreme it may match *x* ensemble shapes, but only one feature on each of them. We noticed that false positives (inactive molecules that meet the threshold) often tend to match fewer features per ensemble shape than the true positives (active molecules that meet the threshold).

This observation can be used to further improve the model by insisting that a threshold number of the ensemble hypotheses be matched per ensemble shape when matching. If this above condition is not satisfied all the bits corresponding to that ensemble shape in the compound signature are neglected. This reflects the intuitive idea that a compound, to be active, not only must fit to a certain shape but also has to make multiple, simultaneous chemical interactions with the active site.

4. SHAPE ALIGNMENT ASSESSMENT

The quality of the model for activity based on shape-feature signatures depends substantially on the ability to align shapes to each other correctly. To assess this ability we conducted a computational experiment on 10 ligands that bind to the thrombin receptor whose three-dimensional X-ray structures in the bound state are available from the Protein Data Bank.[23] The poses for the ligands were standardized by aligning the heavy atoms of the protein structures to within 1.0 Å. This resulted in an alignment of the 10 ligands to each other that we consider to be the *standard* for comparison. The experiment consists of obtaining alignments using our method between all possible pairs of these ligands and comparing the results to the *standard* poses mentioned above. The shape of each bound ligand conformation, l_i, was entered into a shape catalog (i.e. a separate catalog for each ligand). The remaining ligands were then compared to this shape in the catalog. Multiple alignments may be obtained between a pair of ligands (l_i, l_j) because (1) the shape of ligand l_i in the catalog actually consists of all its records (i.e. multiple orientations of the shape) and (2) several orientations (permutations) of the ligand l_j are tried during the matching process. All the alignments obtained for ligand l_j to ligand l_i were compared to the standard pose for l_j by measuring the root-mean-squared distance (RMSD) between the heavy atoms. In Tables 2 and 3 the best root-mean-squared distances obtained for each of these alignments are shown for the two types of alignment procedures (key-feature and centroid) used in our method. Each *(i,j)* entry in this table gives the best root-mean-squared distance obtained when the ligand l_j was compared to the ligand l_i (whose shape was entered into the shape catalog). Not surprisingly, the key-feature alignment, on average, performs better than the centroid alignment in reproducing the standard alignments. All but two pairs had alignments that have an RMSD of less than 2.0 Å. The centroid alignment procedure fails in some cases (∼20%) to provide any alignments. However, with ∼80% of the pairs, it obtained alignments with RMSD values of less than 2.0 Å.

Table 2. Results from Pairwise Alignment of Bound Crystal Ligands Using Our Method with the Key-Feature (Positive Charge) Alignment Procedure, as Compared to the Alignments Obtained from the Protein Structures[a]

	1a3b	1a4w	1a61	1abj	1dwc	1dwd	1etr	1ett	1lhc	1ppb
1a3b	0	1.95	1.26	0.81	1.86	1.8	1.95	1.57	0.63	0.78
1a4w	1.76	0	1.18	1.53	0.54	1.53	0.57	1.26	1.37	1.41
1a61	1.21	1.07	0	0.36	1.23	1.16	1.21	1.35	1.01	0.25
1abj	0.81	2.37	0.36	0	1.94	2.02	1.96	1.91	0.39	0.15
1dwc	1.15	0.64	1.32	0.87	0	0.93	0.28	0.9	0.67	0.78
1dwd	0.48	1.48	0.86	0.71	0.87	0	0.92	0.86	0.48	0.62
1etr	1.22	0.55	1.16	0.84	0.27	0.95	0	1.13	0.67	0.75
1ett	0.92	1.3	1.43	1.03	0.89	0.89	1.11	0	0.95	1
1lhc	0.64	1.67	1.03	0.37	1.67	1.81	1.71	1.67	0	0.38
1ppb	0.78	2.27	0.23	0.15	1.9	1.98	1.92	1.88	0.39	0

[a] The row and column headings give the protein structure IDs from the Protein Data Bank. Each *(i,j)* entry in the table gives the best root mean squared distance obtained when the ligand l_j was compared to the ligand l_i.

Table 3. Results from Pairwise Alignment of Bound Crystal Ligands Using Our Method with the Centroid Alignment Procedure, as Compared to the Alignments Obtained from the Protein Structures[a]

	1a3b	1a4w	1a61	1abj	1dwc	1dwd	1etr	1ett	1lhc	1ppb
1a3b	0	NA	1.2	0.7	1.21	0.67	1.28	1.04	0.28	0.55
1a4w	NA	0	NA	NA	NA	NA	NA	NA	NA	NA
1a61	1.15	NA	0	0.86	NA	NA	NA	NA	1.05	1.08
1abj	0.68	NA	0.85	0	10	1.1	NA	1.25	0.56	0.33
1dwc	1.31	NA	NA	1.5	0	1.13	0.27	0.72	1.3	1.29
1dwd	0.71	NA	NA	1.09	1.12	0	1.1	1.27	0.72	0.98
1etr	1.31	NA	NA	1.54	0.3	1.11	0	0.79	1.31	1.3
1ett	1.13	NA	NA	1.28	0.76	1.3	0.82	0	1.09	1.08
1lhc	0.28	NA	1.03	0.58	1.22	0.7	1.29	1.02	0	0.41
1ppb	0.52	NA	1.1	0.33	1.27	0.98	1.31	1.06	0.38	0

[a] NA means an alignment was not obtained from the method. The row and column headings give the protein structure IDs from the Protein Data Bank. Each *(i,j)* entry in the table gives the best root mean squared distance obtained when the ligand l_j was compared to the ligand l_i.

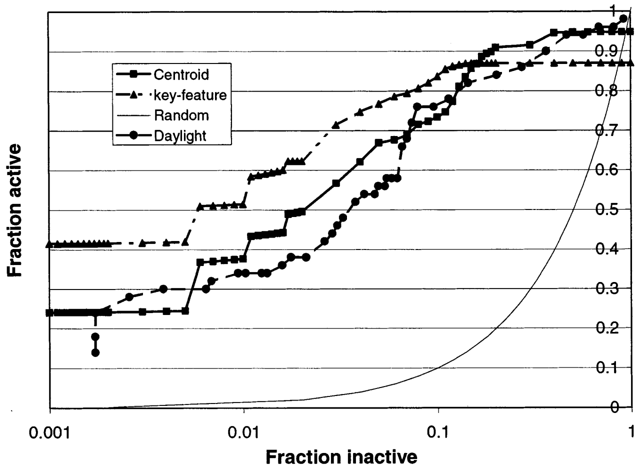

Figure 7. Each of the points on a curve above corresponds to a specific score in the ensemble. It gives the fraction of actives (*y*-axis) and the fraction of inactives (*x*-axis) selected by the model from the test set of compounds. Using the positive charge as the key-feature for alignment in our method, we obtain better enrichment than by using centroid alignment. Both procedures, however, provide better enrichment curves than when using DAYLIGHT fingerprints. Also shown in the plot is a curve for random selection of molecules.

5. RESULTS

This section presents an example application of the use of shape-feature descriptors for screening virtual libraries for biologically active compounds. The results presented here are meant to illustrate the utility of shape-feature descriptors; a more thorough evaluation of the performance of the method has been published in a companion paper.[15]

For the application presented here, a virtual library was constructed and mined for compounds that are inhibitors of thrombin. The virtual library consisted of 1909 compounds, of which 40 were known actives against thrombin. The active molecules were taken from the literature[24,25] and had activities in the micromolar to subnanomolar range. The remaining molecules, the inactive set, were chosen from a chemically diverse internal screening library. Because almost all thrombin inhibitors are positively charged, each molecule chosen for the inactive set was required to have a moiety that would have net positive charge at normal pH.

For this set of compounds, the generation of the model, including the conformational analysis of the molecules, took approximately 20 h (or ~38 s/compound) on a typical desktop computer with 500 MHz Intel Pentium processor and 500MB of RAM. Scoring the compounds in the virtual library was faster because only the ensemble shapes (i.e., those identified as most important for activity) are considered in this stage. A typical run time for scoring compounds is approximately 2 s/compound.

The results presented in Figures 7 and 8 were obtained by performing 5-fold cross-validation runs (i.e., five separate runs). In each cross-validation run, 25% of the data was randomly chosen and withheld as the *test set*. The remaining data (75%) were then used to obtain the *ensemble model*, which consisted of the 50 top ranking bits (hypotheses). Between any pair of ensembles produced from the cross-validation runs, on average 30% of the bits are common. The compounds in the test set were then scored with the ensemble and ranked according to their scores. The performance was then measured by using the following standard technique: For each possible score (i.e., 1−50), the fraction of active (f_a) and fraction of inactive (f_i) compounds in the test set that exceed or equal the score are determined. The value pairs (f_i, f_a) are then plotted in a 2D *performance plot* with the fraction of inactives on the *x*-axis and the fraction of actives on the *y*-axis. For the optimal model, all actives will pass the ensemble model, while no inactives will, which results in a point in the upper left-hand corner of the graph. The random selection of compounds, shown as a continuous line, provides a performance baseline: actives and inactives are selected at the rate at which they are present in the virtual library. Since the number of inactives typically is far larger than the number of actives, performance of the model is critical in the regions where the fraction of inactives selected is small (i.e., low number of false positives). To aid analysis, we expand this portion of the performance plot by using a logarithmic scale on the *x*-axis. Figures 7 and 8 show the performance for two types of activity models constructed using shape-feature descriptors: Figure 7 shows the performance of the shape-feature ensemble model using two different kinds of shape alignments, and Figure 8 shows the performance when the ensemble model is restricted to contain only shapes for which several features are recognized as significant.

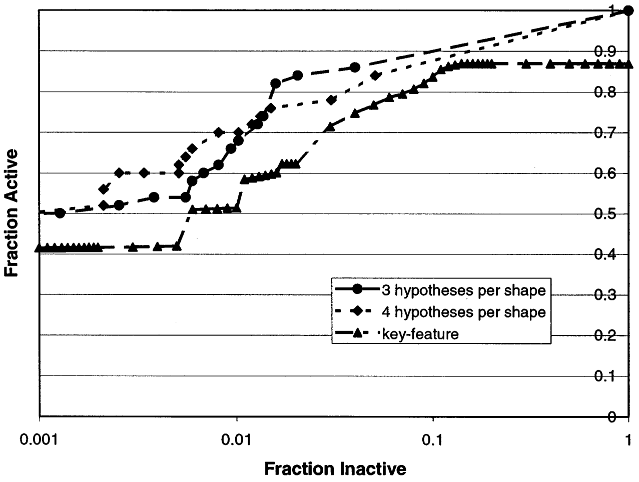

Figure 8. Each of the points on a curve above corresponds to a specific score in the ensemble. It gives the fraction of actives (*y*-axis) and the fraction of inactives (*x*-axis) selected by the model from the test set of compounds. The performance obtained by using key-feature alignment can be improved further by using the multiple hypotheses per ensemble shape restriction. Shown in the plot are the enrichment curves when we insisted that at least 3 or 4 hypotheses per each matched ensemble shape be turned on for a compound.

5.1. Key-Feature and Centroid Alignment. The plot in Figure 7 shows the comparison between the key-feature alignment and the centroid alignment. The multiple hypotheses per shape restrictions described in section 3.4.1 were not applied to the results shown in this plot. Also shown in the plot for reference is the random selection curve and the 2D similarity score curve. The 2D similarity score curve was determined by computing the DAYLIGHT[26] fingerprints, with 2048 bits in each fingerprint. These fingerprints were then processed to generate the model in exactly the same way as the shape-feature signatures. The centroid alignment procedure performs better than the 2D similarity method, in the region of between 0.5% of inactives and 6% of inactive. Further, using key-feature alignment provides better performance than the centroid alignment and the 2D similarity methods.

5.2. Multiple Hypotheses per Shape. Figure 8 shows that the ensemble shape-feature model can be improved further by applying the multiple-hypos per shape restriction described in section 3.4.1. Insisting that each element of the ensemble share the same shape with two or three other elements in the ensemble results in better enrichment curves.

This result is interesting because it suggests that feature combinations within a single shape yield more relevant information for activity than shapes with only one or two significant feature locations. Such combinations can be thought of as pharmacophore models that combine the shape and feature information that is necessary for binding.

6. DISCUSSION

Our method attempts to identify shapes and features that are common among actives and uncommon among inactives by seeking the bits in the shape-feature signatures that relate to activity in a statistically significant manner. The success of the method depends on several factors. The input data must consist of a sufficient number of actives to find steric similarities among them as well as enough inactives to remove those shapes that are easily matched by molecules in general. We have found that we can achieve useful shape-feature models for activity when we have >10 actives and inactives. A second factor is the quality of the conformational analysis used to generate the low-energy conformations used in the model. We know from cases in which the X-ray structure of the bound molecule is known that the conformational model should provide a conformer that is within 1−2 Å root-mean-squared distance from the bound conformation. Thus, a good sampling of the conformational space available to the molecules is essential. Once conformers for the molecules have been generated, they need to be properly aligned, and this is also a crucial step in the method. Alignment of molecules by their centroids was shown to produce good results, but alignment to a unique (key) feature that is common to the active molecules is significantly better. This fact is highlighted by the compounds that inhibit thrombin. Though both benzamidine and NAPAP are thrombin inhibitors, alignment of their centroids shifts the positive charges in the two molecules away from each other, while we know from X-ray data that the charges should be close to coincident. In this case, alignment on this feature is more accurate. Of course, one is not always fortunate enough to have a unique key-feature in each active molecule on which to base the structural alignment.

Our grid based scoring method described in section 3.3 is similar in some ways to CoMFA.[27] However, there are some significant distinctions which deserve mention. The problem of structural alignment of molecules is treated differently in the two methods. CoMFA requires an accurate alignment of conformers as input, and so requires the identification of the relevant conformation for each active molecule. In contrast, the shape-feature method handles an ensemble of conformations for each molecule. Thus, we do not have to identify the important conformation, though we need to make sure it is represented in the ensemble. The method then implicitly identifies the conformation that is interesting. Another difference is the grid representation of the molecules. CoMFA encodes the compounds by real-valued field contributions to each grid point, while we use bits to represent the presence or absence of features at specific locations in shapes. Finally, unlike CoMFA, our method does not attempt to determine real-valued activities but only to classify compounds as those likely to be active and those likely to be inactive. Such classification problems are generally regarded as easier to solve than the calculation of a regression function for activity.[28] Research in scoring functions for docking[29,30] shows that even when structural data are available, the calculation of an accurate regression function for activity is a very difficult problem. Other related approaches are 4D-QSAR,[31] GRIND[32] and, probably closest to our approach, the work by Hahn.[33] The latter approach is similar in terms of the alignment technology. However, it is lacking the analysis part and the classification methodology.

Our method obtains a whole molecule model, i.e., the ensemble shapes belong to the whole ligand conformations, and it may be possible to improve upon this approach by considering the shape of molecular fragments. If one could distinguish the parts of the active molecule that interact with the protein from the parts that are solvent exposed, a better model could result. We are currently investigating using molecular subshapes to build models for bioactivity using much of the same computational methods described here.

7. CONCLUSIONS

We presented a shape-feature based molecular descriptor that can be used to generate models that distinguish active from inactive molecules. Our method does not rely on any structural information about the pharmacological receptor and is fast enough to search large virtual libraries. Cross-validation experiments show that the shape-feature descriptors can produce models for virtual library screening that show significant enrichment and that outperform similar models built using the two-dimensional structural fingerprints generated by the DAYLIGHT software. Further, we found that when a unique chemical feature is common to the set of active molecules and is used as a basis for the structural alignment of shapes, the performance of the resultant activity model improves significantly. We also found that restricting the model to consist of shapes that contain more than two relevant feature positions yields improved bioactivity models. Such a combination suggests that this methodology is able to generate novel pharmacophores that combine chemical features within a steric volume.

REFERENCES AND NOTES

(1) Taylor, R.; Kennard, O. Hydrogen-Bond Geometry in Organic Crystals. *Acc. Chem. Res.* **1984**, *17*, 320−326.

(2) Mills, J. E. J.; Dean, P. M. Three-dimensional hydrogen-bond geometry and probability information from a crystal survey. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 607−622.

(3) Lawrence, M. C.; Colman, P. M. Shape complementarity at protein/protein interfaces. *J. Mol. Bio.* **1993**, *234*, 946−950.

(4) Jones, S.; Thornton, J. M. Principles of protein−protein interactions. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 13−20.

(5) 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.

(6) Babine, R. E.; Bender, S. L. Molecular recognition of protein−ligand complexes: applications to drug design. *Chem. Rev.* **1997**, *97*, 1359−1472.

(7) *DockIt*; Metaphorics: Santa Fe, NM.

(8) Brady, G. P., Jr. Prediction of ligand-binding modes via energy-based genetic algorithm docking. *Book of Abstracts, 218th ACS National Meeting, New Orleans* 1999; COMP-067.

(9) DesJarlais, R. L.; Sheridan, R. P.; Dixon, J. S.; Kuntz, I. D.; Venkataraghavan, R. Docking flexible ligands to macromolecular receptors by molecular shape. *J. Med. Chem.* **1986**, *29*, 2149.

(10) Makino, S.; Ewing, T. J.; Kuntz, I. D. DREAM++: flexible docking program for virtual combinatorial libraries. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 513−532.

(11) Joseph-McCarthy, D.; Thomas, B. E., III.; Alvarez, J. C. Pharmacophore-based molecular docking. *Book of Abstracts, 221st ACS National Meeting, San Diego* 2001; CINF-041.

(12) Kenan, D. J.; Tsai, D. E.; Keene, J. D. Exploring molecular diversity with combinatorial shape libraries. *Trends Biochem. Sci.* **1994**, *19*, 57−64.

(13) Van Drie, J. H. ‘Shrink-wrap’ surfaces: A new method for incorporating shape into pharmacophoric 3D database searching. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 38.

(14) Cramer, R. D.; Poss, M. A.; Hermsmeier, M. A.; Caulfield, T. J.; Kowala, M. C.; Valentine, M. T. Prospective identification of biologically active structures by topomer shape similarity searching. *J. Med. Chem.* **1999**, *42*, 3919−3933.

(15) Srinivasan, J.; Castellino, A.; Bradley, E. K.; Eksterowicz, J.; Grootenhuis, P. D. J.; Putta, S.; Stanton, R. V. Evaluation of a Novel Shape-Based Computational Filter for Lead Evolution: Application to Thrombin Inhibitors. *J. Med. Chem.* **2002**, *45*(12), 2494−2500.

(16) Smellie, A.; Stanton, R. V.; Henne, R.; Tieg, S. Confomational Analysis by Intersection. Submitted to *J. Comput. Chem.* **2001**.

(17) Smellie, A.; Teig, S. L. Conformational analysis by intersection: Ring conformation. *Proceedings of the 217th National Meeting of the American Chemical Society*; Anaheim, 1999.

(18) Foley, J. D.; Van Dam, A.; Feiner, S. K. *Introduction to Computer Graphics*; Addison-Wesley Pub. Co.: 1993.

(19) Conway, J. H.; Sloane, J. A. *Sphere Packings, Lattices and Groups*, 3rd ed.; Springer-Verlag: New York, 1998.

(20) *SMARTS Toolkit;* Daylight Chemical Information Systems: Santa Fe, NM.

(21) Greene, J.; Kahn, S.; Savoj, H.; Sprague, P.; Teig, S. Chemical Functional Queries for 3D Database Search. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1297−1308.

(22) Lanctot, K.; Lemmen, C.; Putta, S.; Greene, J. Using Ensembles to Classify Compounds for Drug Discovery. Manuscript in preparation.

(23) Berman, H. M.; Westbrook, J. Z.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, 235−242, 28 pp.

(24) Vacca, J. P. Thrombosis and Coagulation. *Ann. Rep. Med. Chem.* **1998**, *33*, 81−90.

(25) Malley, M. F.; Tabernero, L.; Chang, C. Y.; Ohringer, S. L.; Roberts, D. G. M.; Das, J.; Sack, J. S. Crystallographic determination of the structures of human a-thrombin complexed with BMS-186282 and BMS-189090. *Prot. Sci.* **1996**, *5*, 221−228.

(26) Daylight Chemical Information Systems: Santa Fe, NM.

(27) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(28) Vapnik, V. N. *Statistical Learning Theory*; John Wiley & Sons: New York, 1998.

(29) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based Scoring Function to Predict Protein−Ligand Interactions. *J. Mol. Biol.* **2000**, *295*, 337−356.

(30) Muegge, I.; Martin, Y. C. A general and fast scoring function for protein−ligand interactions: a simplified potential approach. *J. Med. Chem.* **1999**, *42*, 791−804.

(31) Venkatarangan, P.; Reaka, A.; Lakshamanan, U.; Duca, J. S.; Hopfinger, A. J. 4D-QSAR analysis of a set of glucose analogue inhibitors to glycogen phosphorylase. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1141−1150.

(32) Pastor, M.; Cruciani, G.; McLay, I.; Pickett, S.; Clementi, S. Grid Independent Descriptors (GRIND). A Novel Class of Alignment-Independent Three-Dimensional Molecular Descriptors. *J. Med. Chem.* **2000**, *43*, 3233−3243.

(33) Hahn, M. Three-Dimensional Shape-Based Searching of Conformationally Flexible Compounds. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 80−86.

CI0255026