Shape Signatures: A New Approach to Computer-Aided Ligand- and Receptor-Based Drug Design

Randy J. Zauhar,*,[†] Guillermo Moyna,[†] LiFeng Tian,[†] ZhiJian Li,[†] and William J. Welsh[‡]

Department of Chemistry & Biochemistry, University of the Sciences in Philadelphia, 600 S. 43rd Street, Philadelphia, Pennsylvania 19104 and Department of Pharmacology, Robert Wood Johnson Medical School, University of Medicine and Dentistry of New Jersey, 675 Hoes Lane, Piscataway, New Jersey 08854

Received May 20, 2003

A unifying principle of rational drug design is the use of either shape similarity or complementarity to identify compounds expected to be active against a given target. Shape similarity is the underlying foundation of ligand-based methods, which seek compounds with structure similar to known actives, while shape complementarity is the basis of most receptor-based design, where the goal is to identify compounds complementary in shape to a given receptor. These approaches can be extended to include molecular descriptors in addition to shape, such as lipophilicity or electrostatic potential. Here we introduce a new technique, which we call *shape signatures*, for describing the shape of ligand molecules and of receptor sites. The method uses a technique akin to ray-tracing to explore the volume enclosed by a ligand molecule, or the volume exterior to the active site of a protein. Probability distributions are derived from the ray-trace, and can be based solely on the geometry of the reflecting ray, or may include joint dependence on properties, such as the molecular electrostatic potential, computed over the surface. Our shape signatures are just these probability distributions, stored as histograms. They converge rapidly with the length of the ray-trace, are independent of molecular orientation, and can be compared quickly using simple metrics. Shape signatures can be used to test for both shape similarity between compounds and for shape complementarity between compounds and receptors and thus can be applied to problems in both ligand- and receptor-based molecular design. We present results for comparisons between small molecules of biological interest and the NCI Database using shape signatures under two different metrics. Our results show that the method can reliably extract compounds of shape (and polarity) similar to the query molecules. We also present initial results for a receptor-based strategy using shape signatures, with application to the design of new inhibitors predicted to be active against HIV protease.

Introduction

A universal problem in computer-aided drug design is the comparison of molecular shape.[1−3] In ligand-based design, the underlying assumption is that a biologically active compound is complementary in shape to some target receptor, suggesting that molecules similar in shape and electrostatic properties to a known active compound will themselves be complementary to the receptor and also active. In receptor-based design, the structure of the target binding site is already known in atomic detail, and the goal is to directly identify compounds that are complementary to the site both in shape and polarity.

A number of methods have been devised for screening compound libraries for molecules likely to be active against a selected target.[4−12] Most of these take molecular shape into account, either explicitly or implicitly. Perhaps the most popular ligand-based strategy that takes shape explicitly into account is CoMFA[13,14] (comparative molecular field analysis) wherein the van der Waals and electrostatic fields of molecules are sampled over a grid and used as descriptors in a regression model intended to predict biological activity. CoMFA thus includes both molecular shape and polarity. The various methods for defining *pharmacophore* models represent ligand shape implicitly by incorporating some collection of hydrogen bond acceptors and donors and regions of steric bulk and imposing intergroup distance constraints; this 3D geometric information clearly depends on molecular shape. A number of approaches have been developed that compute *topological descriptors* of molecules, beginning with chemical structure or starting with the wave function; such descriptors derive directly from molecular shape. Even methods based on *chemical fingerprints* include implicit shape information, since only a restricted family of compounds will be compatible with the chemical and connectivity information contained in the fingerprint.

Receptor-based design strategies generally involve an explicit representation of shape derived from an atomic-resolution structure of the active site. For example, UCSF DOCK[15,16] packs the active site with spheres, producing an efficient representation of the volume available to accommodate a ligand and combines this with positions of hydrogen bond acceptors and donors. Docking algorithms such as FLOG,[17] GOLD,[18,19] and FlexiDock[20] use an all-atom representation of the active

* To whom correspondence should be addressed. Phone: 215-596-8691, Fax: 215-596-8543, e-mail: r.zauhar@usip.edu.

[†] University of the Sciences in Philadelphia.

[‡] University of Medicine and Dentistry of New Jersey.

10.1021/jm030242k CCC: $25.00
Published on Web 11/19/2003

site and thus represent its geometry in fine detail. Pharmacophore-like models can also be devised for receptors, and these include shape information in the same way as ligand-based models.

It is clear that shape information is key to the identification of molecules that are likely to be biologically active. At the same time, this information is difficult to efficiently encode and to use in database searching, in either ligand- or receptor-based design. CoMFA and pharmacophore methods use a large amount of explicit information to encode shape, involving many grid points or geometric constraints. Scanning a chemical library with a pharmacophore query involves much computation, since each molecule must be repositioned and flexed in order to determine if it can fit the model. Similarly, receptor-based docking strategies require many packed spheres and/or atom positions to encode the shape of the active site, and again scanning a chemical library with a docking program requires many detailed calculations for each compound considered, often involving molecular mechanics computations, or at the very least “bump checks” to test shape compatibility between receptor and ligand.

While improved efficiency in pharmacophore search and docking methods has made such approaches usable for screening chemical libraries, there is clearly no upper limit on the number of compounds that we would like to be able to consider. A method for rapidly comparing shape constitutes a desirable addition to the computational arsenal for selecting active molecules. A method for screening libraries by shape may by itself yield compounds of great interest, and in addition such compounds may then be passed down a “computing pipeline” for additional screening (e.g. computation of log *P*) or be used in detailed docking studies.

In this article we describe a novel approach we call *shape signatures* for compactly representing molecular shape and demonstrate how the method can be easily applied to both ligand- and receptor-based molecular design. Our approach uses *ray-tracing* to explore the volume interior to a ligand or the space exterior to a receptor site. Shape signatures are probability distributions derived from the aforementioned ray-traces, and they serve as compact descriptors of shape requiring modest storage space and which may be quickly compared to test for shape similarity or complementarity. While augmenting a chemical compound library with shape signatures requires a significant computational expense, this price need by paid only once, when the signature component is added to the database; moreover, these calculations, while significant, are tractable and can be carried out using readily available computing facilities.

We will discuss both the initial implementation of shape signatures, which includes only shape information, as well as extensions which couple the existing ray-tracing technique with other computed molecular properties to define signatures with higher-dimensional domains. Specifically, we will illustrate an approach that includes the molecular electrostatic potential (MEP) to define a two-dimensional (2D) signature. These 2D-MEP-based signatures combine shape and polarity information and can be used to select molecules that are similar in shape and electrostatic potential to a query.

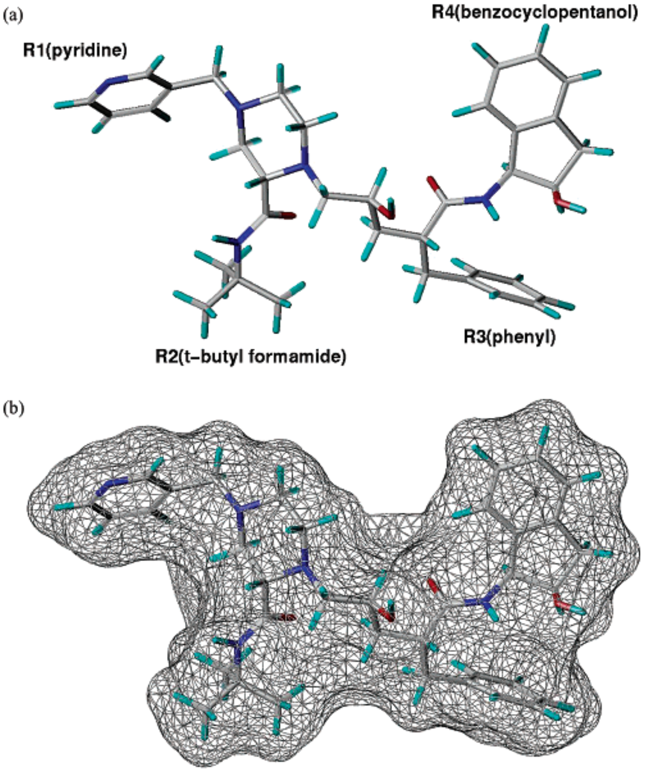


Figure 1. (a) Structure of Indinavir, including substituent terminology used here. (b) Indinavir with triangulated solvent-accessible surface generated using SMART. The Indinavir structure was taken from PDB entry 1HSG,[34] which was also the source of the atomic coordinates for native HIV protease used in this work.

Methods

Ray-tracing. In the shape signatures approach, the shape of a molecule is assumed to coincide with its solvent-accessible molecular surface,[21−23] which is generated in the usual way by the points of contact of a rolling spherical probe. In our application, we need a detailed representation of the surface, which is best realized by breaking the surface into small area elements. To accomplish this we use the smooth molecular surface triangulator algorithm (SMART)[24] which has been described previously. SMART partitions the molecular surface into regular triangular area elements, which are well-suited to the computations that follow. The definition of the solvent-accessible molecular surface depends on the choices of atomic radii, solvent probe radius, and the density of element corners (vertexes) to be generated. In this work, we use the PARSE[25] atomic radii, a radius for the solvent probe of 1.4 Å, and vertexes spaced approximately 0.5 Å apart. Figure 1a shows the chemical structure of the HIV protease inhibitor Indinavir,[34] and Figure 1b shows a triangulated molecular surface for this molecule.

The volume defined by the molecular surface is explored using a modified form of ray-tracing, a technique widely used in presentation graphics and computer animation. In graphics-oriented applications, ray-tracing means tracking the paths of light rays that emanate from some number of defined sources and

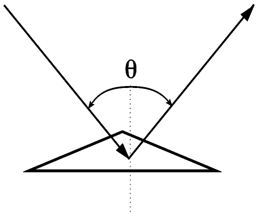


Figure 2. Geometry of ray-tracing. Here an incoming ray is reflected by a single triangular element which forms part of the molecular solvent-accessible surface. The component of the incoming ray parallel to the plane is unchanged by reflection, while the component perpendicular to the plane is reversed.

which are then reflected by objects in a scene. In its full realization, ray-tracing takes into account the material properties of objects as well as the atmosphere the rays travel through. For our purposes of describing shape, the requirements are simpler; we need only consider "perfect" reflection from the molecular surface, as illustrated in Figure 2. Furthermore, we have no light sources, but rather start each ray from a randomly selected point on the molecular surface and then allow the ray to propagate by the rules of optical reflection.

In our approach, a ray is initiated at the midpoint of a triangular surface element chosen at random, with initial direction defined by selecting a second point at random in a hemisphere centered at the midpoint of the planar element. If we are generating a ray-trace for a ligand or other small molecule, then the hemisphere lies on the interior side of the element as determined by the outward-facing surface normal (which is defined by the SMART algorithm). If on the other hand we wish to define the shape of a receptor site, then the hemisphere lies on the *outward side* of the element, and the initial ray propagation is directed toward the exterior of the molecule. When performing such an exterior ray trace, the user supplies a list of atoms that define the receptor site of interest, and only those surface elements that are close to the site atoms are involved in ray propagation, either as initiation points for new rays or as reflection points.

Once a ray is initiated, it is propagated by the rules of optical reflection, and reflection points are written to a file. (The user specifies the number of reflections to generate.) Three events may terminate the propagation of a ray: (1) the number of reflections equals the number requested by the user; (2) the propagating ray makes a "glancing" contact with a surface element or strikes too close to the boundary between two adjacent elements, leading to mathematical difficulties in computing the reflection angle, or (3) the ray strikes no surface element and heads out to infinity—this is only possible in exterior ray-traces of receptor sites. In case 1, the algorithm is finished. In cases 2 and 3, the ray-trace is simply restarted at a newly chosen point on the molecular surface. Figure 3 shows two ray-traces for Indinavir, with 100 and 10000 reflection points.

Shape Signatures. A ray-trace provides raw information about three-dimensional shape but is not useful in itself. Our aim is to derive *probability distributions* that characterize the ray-trace and to use these as compact descriptors of shape (and, as we will later demonstrate, molecular polarity as well). Shape signatures are nothing but these ray-trace-derived probability distributions. In this work, we will represent our probability distributions as histograms, but we recognize that other representations (e.g. wavelets) are available and may ultimately prove to be more useful.

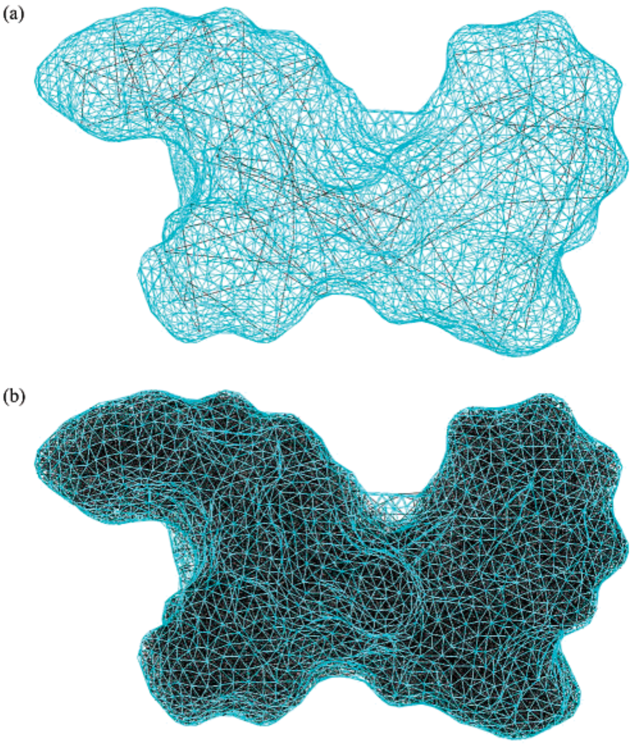


Figure 3. Ray-traces for Indinavir. (a) Low density (100 reflections). (b) High density (10000 reflections). Most of the volume is densely filled, with the exception of two small regions. In a Connolly representation of the surface these regions would represent self-intersecting solvent-accessible surface and would be removed from the triangulation; in the SMART representation they are retained but are narrow and of very small volume.

We term the line segment that connects two successive reflection points a *ray-trace segment*. Perhaps the simplest shape signature is the distribution of the lengths of these segments. We call this a *one-dimensional (1D) signature* to emphasize that the domain of the probability distribution (namely segment length) has one dimension. Figure 4 shows the distribution of segment lengths for Indinavir, derived from 10000 and 50000-point ray-traces. It is observed that signatures converge rapidly with increasing number of reflections and are not sensitive to the initiation point of the ray-trace.

We also define signatures with higher-dimensional domains, which incorporate additional molecular descriptors. One approach to generating two-dimensional signatures is to associate a *surface property*, measured at each reflection point, with the sum of the segment lengths on either side of the reflection. An obvious and important property is the molecular electrostatic potential (MEP) computed over the molecular surface. Figure 5a illustrates this approach to defining a two-dimensional domain, and Figure 5b shows 2D MEP-based signatures for Indinavir using 10000 and 50000 reflections. 2D-MEP signatures are joint probability distributions for observing a sum of segment lengths together with a particular value of the electrostatic

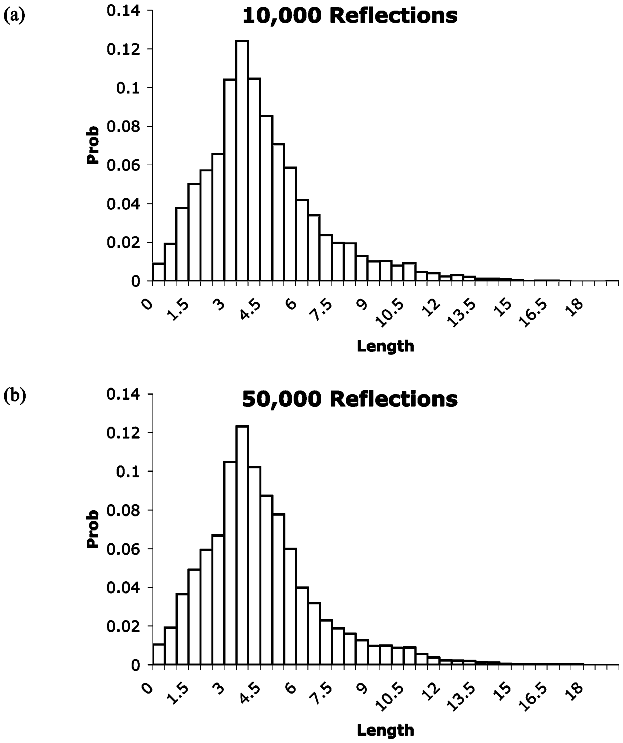


Figure 4. 1D shape signatures for Indinavir, for (a) 10000 and (b) 50000 reflections. Integrated difference between the two distributions is 0.037 probability unit. Difference between 20000 reflection distribution (not show) and 50000 reflection distribution (b) is 0.023.

potential at a given reflection point. They thus simultaneously encode information concerning shape and polarity.

Shape signatures are clearly independent of molecular orientation and furthermore involve no overlay of a grid on the molecule (as in CoMFA), with accompanying questions as to the effects of grid spacing and orientation on results. These constitute key advantages of our approach.

Shape Signature Comparison. Clearly, shape signatures are useful only in being compared. In brief, we wish to use shape signatures to rapidly test for shape *similarity* between one molecule and another, and shape *complementarity* between a molecule and a protein receptor site, representing ligand- and receptor-based strategies.

We compare signatures by measuring the distance between the associated histograms, using simple metrics that can be computed quickly. The most elementary metric is based on the L_1 norm commonly used to compare functions (eq 1):

$$L_1 = \sum_i |H_i^1 - H_i^2| \tag{1}$$

Here the subscript i ranges over the union of all the bins for histograms H^1 and H^2 (it is assumed throughout that the bins for any two histograms will have the same alignment and so fall into a simple one-one correspondence). We assume that our probability distributions are normalized, so that the sum of the histogram heights over all the bins is unity; then under the L_1 metric the *maximum* distance between two histograms is 2, in which case the histograms being compared have no common support (i.e. no bin positions where both functions simultaneously have nonzero height). The *minimum* distance between two histograms, under this or any other acceptable distance measure, is zero (corresponding to the case where the distributions being compared are identical).

It has been observed that histograms often feature a dominant peak around 3 Å, which clearly arises from ray-trace segments that "measure" small-scale as opposed to large-scale molecular shape. In an attempt to amplify the sensitivity to overall molecular shape when making comparisons, we have also used the following modified metric (eq 2):

$$R_1 = \sum_i d_i |H_i^1 - H_i^2| \tag{2}$$

We call this a "ramp" metric since it weights the ith term in the sum by a ramp function (the length d_i associated with the ith bins of the histograms).

A key advantage of the shape signatures approach is that a comparison using either of these metrics requires little computing time, involving arithmetic on histograms that typically have fewer than 50 bins. In this it is comparable to chemical fingerprint methods, which also require few operations to compare two molecules.

The analogues of the preceding metrics for 2D signatures are

$$L_1^{2D} = \sum_i \sum_j |H_{i,j}^1 - H_{i,j}^2| \tag{3}$$

and

$$R_1^{2D} = \sum_i \sum_j d_i |H_{i,j}^1 - H_{i,j}^2| \tag{4}$$

where index i varies over the length indices of the bins, and j varies over the second dimension (which in all the examples here will be an electrostatic potential scale). 2D signatures obviously require more arithmetic operations than 1D signatures to compare, but the computational expense is still limited and much closer to fingerprint comparisons than to strategies that involve docking or reorientation of the molecules being compared.

Segment Culling. We have already noted that 1D signatures typically include a large peak at about 3 Å segment length. Closer examination of the ray-traces from which the signatures are derived reveal a large number of segments that span atomic diameters, as illustrated in Figure 6. Since these would appear to encode relatively little information about the overall shape of the molecule, the ray-tracing program was modified so that segments that involve reflections at the same atom are discarded, and additional segments are generated as needed to match the user-specified total number of reflections. This segment "culling" ensures that the surviving segments encode useful information about the overall shape of the molecule being considered. Since this procedure entails a significant added computational expense, some database searches were performed both with and without segment culling, as described below, to evaluate its merit.

Implementation. Our goal is to begin with a database of molecules, which we assume are saved in Tripos

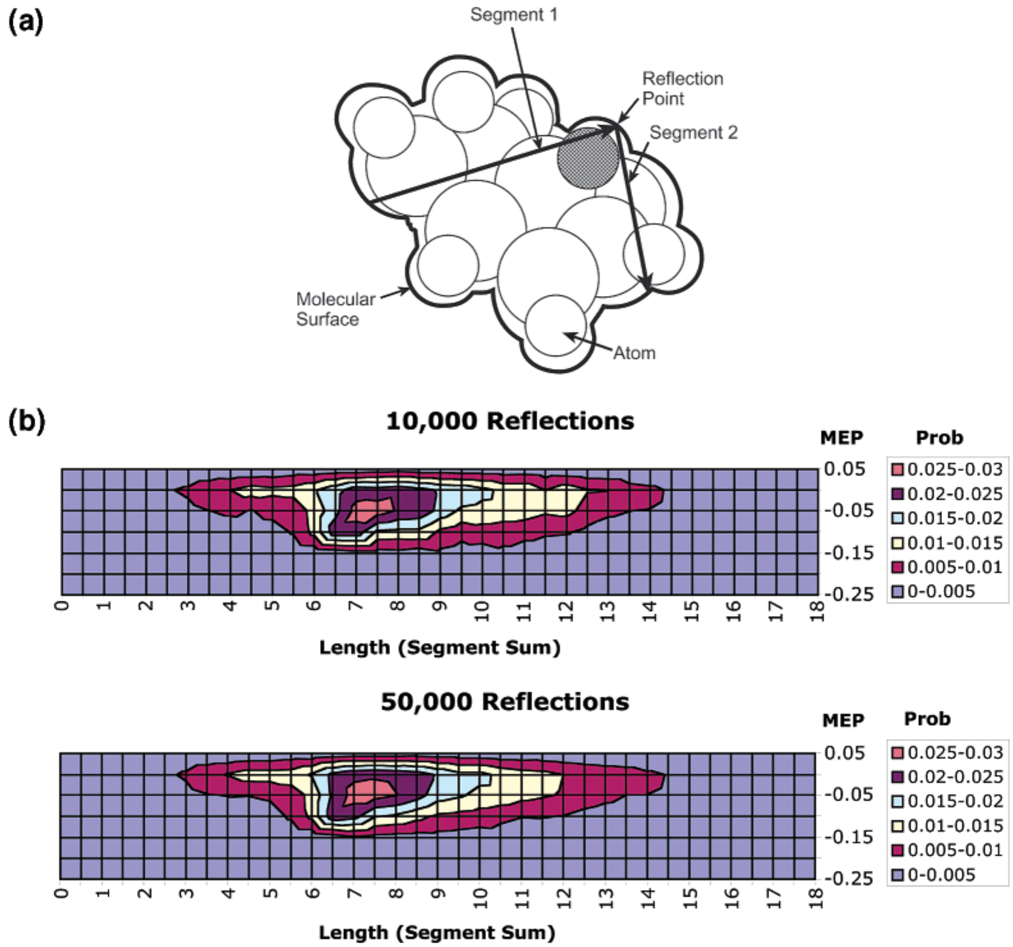


Figure 5. (a) Scheme for defining a two-dimensional shape signatures domain. We compute the joint probability distribution for observing a summed length of segments (1 and 2) on either side of a reflection point, together with the MEP computed at the reflection. (b) 2D-MEP signatures for Indinavir, for 10000 and 50000 reflections.

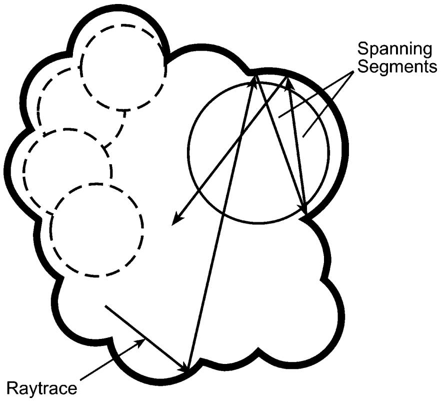


Figure 6. Segment culling. The spheres represent atoms, the heavy outline the solvent-accessible surface Here the ray-trace includes several segments that span the diameter of a single atom. These "uninformative" segments can be eliminated if desired, thereby increasing the proportion of segments that encode information about the overall shape of the molecule.

MOL2 format and from this to generate a database of 1D- and 2D-MEP shape signatures. The molecules must have 3D atomic coordinates and partial atomic charges which we have assigned using the Gasteiger method[26] for the work described here. Processing of the molecular structure database is carried out using a C-shell script which performs the following operations for each molecule:

(1) Generation of a triangulated surface using SMART, which is implemented as a C program.

(2) Ray-tracing within the surface using a second C program. Since this is the most computation-intensive step, efficiency is an important concern. A fast algorithm using a grid acceleration method is employed here. This program creates a file with a specified number of ray-trace segments.

(3) Accumulation of histograms is performed by a third C program that reads the ray-trace file and sums the occurrences of segment lengths, using a bin size specified by the user. The program also computes the electrostatic potential using the partial atomic charges contained in the molecular structure file,

$$\Phi(\mathbf{r}_p) = \sum_j \frac{q_j}{|\mathbf{r}_p - \mathbf{r}_j|} \tag{5}$$

where Φ is the molecular electrostatic potential (MEP) computed at reflection point $\mathbf{r}_p$, and the index j ranges over all the atoms, which have positions $\mathbf{r}_j$ (measured in Å) and partial charges q_j (measured in elementary charge units). The MEP values computed at each reflection point, along with the sum of the segment lengths that adjoin the reflection point, are used to accumulate a 2D-MEP-based signature as illustrated in Figure 5. The resulting histograms are written to an ASCII file, with a format that includes all pertinent information (number of reflections, bin size, etc.)

(4) Finally, a PERL script adds the histogram information generated in operation 3 to a growing database of 1D- and 2D-MEP shape signatures.

In our implementation of shape signatures, the query used to scan a database is itself a database of signatures which could refer to a single object. The query database could be generated from a set of small molecules, in which case the same procedure described above is employed, or a receptor site. In the latter case, the procedure is similar, except that an exterior ray trace is performed (typically over a protein), and the user must specify a set of atoms that define the receptor site. In either case, comparison of the query and target database is effected using a C program that compares each histogram in the query database against all those in the target using one of the metrics described above and writes out a hit list file that reports the best *n* hits for each of the queries (the hit-list length *n* is chosen by the user).

We note that while our metrics can readily compare receptor sites and ligands for shape complementarity, it is less straightforward to measure *electrostatic* complementarity. For example, one might simply reverse the sign of the electrostatic field for either query or target signature and proceed using either of the existing metrics; one then finds an exact match only if query and target have exactly complementary shapes, *and* electrostatic fields that are equal and opposite. This is an extraordinarily stringent criterion, which would lead to poor scores for clearly useful matches. In recognition of this problem, we have not attempted to carry out 2D-MEP searches using receptor-based queries.

Applications

Strategy. Our approach has been to first focus on a small, manageable database of known composition and to assess the efficacy of the shape signatures approach in matching compounds similar in shape and polarity under a variety of options (segment culling enabled or disabled, along with use of either the L_1 or R_1 metric). This provides the opportunity to evaluate the method in a situation where potential hits (and misses) are known beforehand and also to develop a sense for the sizes of 1D and 2D scores that are associated with “interesting” matches. This is a necessary prelude to applying the method to larger and more diverse databases.

Tripos Fragment Database. We initially applied the shape signatures method to the Tripos fragment database, a diverse collection of small molecules including heterocycles, carbohydrates, amino acids, and nucleotides, which is supplied as a standard component of the SYBYL molecular modeling package.[20] This database was especially useful for our initial tests given its small size and its incorporation of multiple representatives of each family of compound (ensuring that a given query from the database will usually have several potential matches). Very small fragments were removed from the database at the start and also some perfectly linear molecules (e.g. allene) which were not handled well by the SMART surface algorithm. This left a total of 235 compounds. Dummy atoms were removed from the amino acids in the database, the resulting empty valences filled with hydrogens, and the side chains of glutamic acid, aspartic acid, lysine, and arginine modified to correspond to the ionized form. Gasteiger charges were assigned to all the compounds in the final set, and 1D- and 2D-MEP signatures were generated using either 50000 or 250000 reflections in each ray-trace in separate computational experiments. The signatures were assembled into databases as described above.

Each resulting database was compared against itself (i.e. each compound in the database was used as a query and compared against all the remaining compounds). The L_1 and R_1 metrics (eqs 1−4) were used in separate comparison. For each query, the 10 best (lowest-scoring) hit compounds were retained. This self-comparison was carried out for both the 50000- and 250000-reflection databases, using 1D- and 2D-MEP signatures, and with segment culling either enabled or disabled. Examination of the hit compounds in the context of their scores were used to propose score cutoffs to distinguish those matches likely to be interesting.

NCI Database Preparation. The National Cancer Institute compound database[27−32] as bundled with the SYBYL UNITY tools was used as a source of molecules for creation of a shape signatures database with 1D- and 2D-MEP signatures. The starting database was screened for all compounds with molecular weight less than 800 Da, yielding 113826 molecules. Gasteiger charges were computed for all of the molecules in the resulting working set. 1D- and 2D-MEP shape signatures were computed for all the compounds, using a sixteen-processor Beowulf cluster. It should be pointed out that each processor was simply allotted a fraction of the molecules to be analyzed, and there was no need to employ the use of parallel code. 50000 reflections were generated in the ray-trace for each compound, and segment culling was employed, as described above. Of the compounds processed, about 0.4% failed (in every case due to an error in molecular surface generation), yielding a total of 113331 compounds in the NCI shape signatures database used in subsequent work. (No attempt at this point has been made to carefully assess the reasons for failure in surface generation or to explore ways to reduce the error rate.) Preparation of the database consumed approximately 100 hours wall-clock time on a sixteen-processor cluster of 450 MHz Pentium-III processors running under the Linux operating system.

Comparison of Tripos Database against NCI. All of the 1D and 2D signatures (50000 reflections per signature) for the Tripos fragment database were used as queries against the NCI shape signatures database described above. The best 50 hits for each query were collected. Searches were carried out using 1D- and 2D-MEP signatures, along with either L_1 or ramp metrics (eqs 1−2, and 3−4, respectively) for a total of four searches. Six query compounds, comprising a set that is both structurally diverse and biologically interesting, were selected for detailed examination here.

A special concern when comparing a query against a large database is the distribution of scores. To be useful, a search method must exhibit a high degree of selectivity, so that truly interesting hits have scores that differ markedly from the mean. In other words, it should be possible to identify a reasonable cutoff score which can be applied to extract a relatively small and meaningful set of hits from a diverse target database. To examine the character of the scores distribution for shape signatures, a special version of the search program for the L_1 metric was prepared which accumulated score statistics in a file. Searches were carried out against the NCI database using this program for all the molecules in the Tripos fragment database, for both 1D- and 2D-MEP signatures. It was thus possible to accumulate the distribution of scores from this relatively large database and to express the number of observed hits as a function of score for each query molecule.

Use of the NCI Database for Receptor-Based Design. We used the HIV protease inhibitor Indinavir as a starting framework. As shown in Figure 1a, the compound includes pyridine (R_1), *tert*-butyl formamide (R_2), phenyl (R_3), and benzocyclopentanol (R_4) as substituents, which are attached to a framework containing piperazine, a peptide group, and a central hydroxyl which marks the site of the transition state analogue presented by the inhibitor.

Rather than attempt to find receptor-based matches to the entire binding site, we took the approach of finding matches to receptor subsites, which we defined by excising these substituents one-at-a-time from the experimental complex of the inhibitor and the native protease molecule.[34] In this way four separate subsites were generated, each marked with a SYBYL dummy atom attached to the portion of the inhibitor that remained. One of these sites, R_1, was largely exposed to solvent and did not provide a well-defined, enclosed pocket; it was omitted from the analysis below, and the original substituent (pyridine) was retained at this position.

Ray-tracing was performed in each pocket with 50000 reflections, and the ray-traces were used to generate 1D histograms for the three sites considered (R_2, R_3, and R_4). These were used to search the NCI database for compounds of shape similar to the pocket volumes (or stated another way, of shape complementary to the receptor subsites). Parameters were identical to those used in the ligand-based searches.

Once a collection of hits was assembled for each subsite, we were faced with the task of attaching these to the framework. This was done using a custom SYBYL application program called ALMS (*A*utomated *L*igand-binding with *M*ultiple *S*ubstitutions[33]) written in the SYBYL programming language (SPL). In brief, each nonring hydrogen of an NCI hit molecule was considered as a possible attachment point, and each ring hydrogen as a possible attachment point through an added methylene carbon. In this way, a single hit molecule was “exploded” into a family of fragments, each with a single free valence, marked by a dummy atom. A fragment was attached to the framework by removing the dummy atoms on both the hit and the target inhibitor site and replacing these with a single bond linking the inhibitor and the fragment. The orientation of the newly attached fragment was then optimized using FlexiDock, the genetic-algorithm-based optimizer included with the SYBYL modeling package. Default force-field settings were used in FlexiDock (including hydrogens with reduced van der Waals radius and epsilon parameter), and in all calculations the genetic algorithm proceeded for 500 generations.

Each framework variable site was considered individually, with additions of all the fragments generated from the best *n* hits for a particular site carried out with the substituents of the starting inhibitor in place at all the other sites. The FlexiDock inhibitor−receptor interaction energies computed after adding all of the fragments targeted to a particular site were used to rank the fragments for that site. Next, the top *k*, *m*, and *n* fragments for sites R_2, R_3, and R_4, respectively, were added in all possible combinations, with precomputed optimized geometry, generating $k \times m \times n$ ligand molecules. A final energy minimization was performed in the field of the frozen receptor for each ligand, followed by an updated computation of the interaction energy. The interaction energies so computed were used to rank the table of putative inhibitors finally generated.

We will discuss ALMS, the system for the fragment-based combinatorial construction of inhibitors used in this study, in more detail elsewhere.

Results

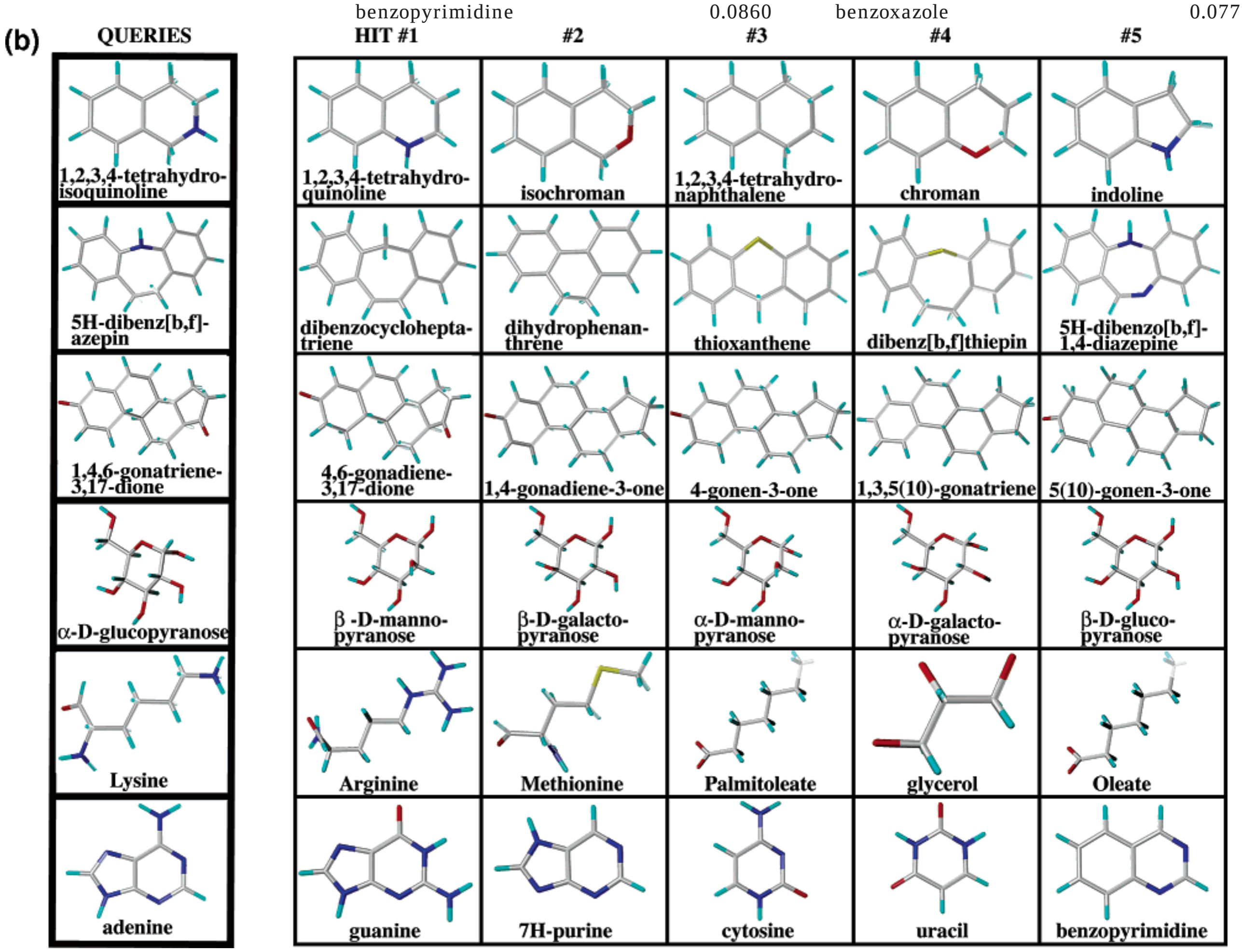
Tripos Fragment Database. Table 1a shows hits found for a selection of query compounds, where we have compared the Tripos fragment database against itself, with 50000 reflections per histogram using 1D signatures and the L_1 metric and with segment culling enabled and disabled. Table 1b shows structures of the top five hits for each of the six queries, for the case of segment culling enabled.

It is seen that the 1D signatures perform well in selecting compounds chemically or structurally similar to the query. This observation is amplified by examining all of the available data for the Tripos database. Moreover, one compound of a class generally selects all compounds of the same class present in the database; for example, a fatty acid (laurate) selects all other fatty acids present in the database, a carbohydrate (α-D-glucopyranose) other carbohydrates, an amino acid (lysine) other amino acids, etc. Dispensing with the segment-culling procedure affects the size of the scores slightly, usually making the distances between histograms a bit smaller but clearly has little effect on the rank order of hits in this example. Switching to the R_1 metric changes the size of the scores as would be expected but has little impact on the rank order of hits for most of the queries. Again, eliminating segment culling does perturb the scores but does not significantly modify the order of hits. Given the similarity of these results to those found under the L_1 metric, they will not be tabulated here. (The authors may be contacted for these or any other results mentioned in this article but not reported in detail.)

When used in the Tripos database self-comparison, 2D-MEP signatures produce results similar to those of the 1D search but with some changes in hit ranking that exhibit sensitivity to the electrostatic properties of query and target compounds and with a much smaller number of meaningful hits. This is not surprising, given that the 2D-MEP searches select simultaneously on the basis of shape and polarity and thus are more stringent than 1D searches. Examples of this are seen in Table 2 where we present results for 2D-MEP signatures compared under the L_1 metric. For example, where the query lysine selected methionine among the top five hits when

Table 1. (a) Results for Six Query Compounds: 1-D Shape Signature Self-Comparison of Tripos Fragment using L_1 Metric

	culling		no culling	
query	hit	score	hit	score
1,2,3,4-tetrahydroisoquinoline	1,2,3,4-tetrahydroquinoline	0.0370	1,2,3,4-tetrahydroquinoline	0.0173
	isochroman	0.0386	isochroman	0.0316
	1,2,3,4-tetrahydronaphthalene	0.0490	chroman	0.0399
	chroman	0.0574	1,2,3,4-tetrahydronaphthalene	0.0475
	indoline	0.0767	indan	0.0525
5*H*-dibenz[*b*,*f*]azepin	dibenzocycloheptatriene	0.0351	dibenzocycloheptatriene	0.0332
	dihydrophenanthrene	0.0482	dihydrophenanthrene	0.0384
	thioxanthene	0.0578	thioxanthene	0.0466
	dibenz[*b*,*f*]thiepin	0.0695	5*H*-dibenzo[*b*,*f*]-1,4-diazepine	0.0487
	5*H*-dibenzo[*b*,*f*]-1,4-diazepine	0.0800	dibenz[b,f]thiepin	0.0578
1,4,6-gonatriene-3,17-dione	4,6-gonadiene-3,17-dione	0.0502	4,6-gonadiene-3,17-dione	0.0400
	1,4-gonadien-3-one	0.0743	1,4-gonadien-3-one	0.0660
	4-gonen-3-one	0.0984	4-gonen-3-one	0.0838
	1,3,5(10)-gonatriene	0.0986	1,3,5(10)-gonatriene	0.0862
	5(10)-gonen-3-one	0.1004	5(10)-gonen-3-one	0.0984
α-D-glucopyranose	*β*-D-mannopyranose	0.0417	α-D-mannopyranose	0.0376
	β-D-galactopyranose	0.0420	*β*-D-mannopyranose	0.0379
	α-D-mannopyranose	0.0559	*β*-D-galactopyranose	0.0391
	α-D-galactopyranose	0.0744	α-D-galactopyranose	0.0560
	β-D-glucopyranose	0.0748	*β*-D-glucopyranose	0.0766
lysine	arginine	0.0862	methionine	0.0527
	methionine	0.1024	arginine	0.0821
	palmitoleate(C16)	0.1163	laurate(C12)	0.0959
	glycerol(-H)	0.1179	palmitoleate(C16)	0.1004
	oleate(C18)	0.1202	myristate(C14)	0.1006
adenine	guanine	0.0626	guanine	0.0388
	7*H*-purine	0.0712	7*H*-purine	0.0701
	cytosine	0.0840	benzimidazole	0.0743
	uracil	0.0854	1*H*-indazole	0.0747
	benzopyrimidine	0.0860	benzoxazole	0.0775



only shape was considered (Table 1), the 2D search results do not include this compound. Our initial experience with this metric suggests that meaningful 1D matches between query and target usually involve

Table 2. Results for Six Query Compounds: 2D-MEP Shape Signature Self-Comparison of Tripos Fragment Database Using L_1 Metric

	culling		no culling	
query	hit	score	hit	score
1,2,3,4-tetrahydroisoquinoline	1,2,3,4-tetrahydroquinoline	0.0847	1,2,3,4-tetrahydroquinoline	0.0762
	1,2,3,4-tetrahydronaphthalene	0.1496	1,2,3,4-tetrahydronaphthalene	0.1307
	indoline	0.1732	indoline	0.1320
	acenaphthene	0.1908	indan	0.1554
	indan	0.2161	acenaphthene	0.1804
5*H*-dibenz[*b*,*f*]azepin	dibenzocycloheptatriene	0.1116	dibenzocycloheptatriene	0.1031
	acridan	0.2089	acridan	0.1538
	5*H*-dibenzo[*b*,*f*]−1,4-diazepine	0.2109	5*H*-dibenzo[*b*,*f*]−1,4-diazepine	0.1672
	1,2,3,4-tetrahydroisoquinoline	0.2268	phenanthridine	0.1762
	1,2,3,4-tetrahydroquinoline	0.2292	dihydrophenanthrene	0.1802
1,4,6-gonatriene-3,17-dione	4,6-gonadiene-3,17-dione	0.0888	4,6-gonadiene-3,17-dione	0.0852
	5a-gonane-3,17-dione	0.1383	5a-gonane-3,17-dione	0.1383
	1,4-gonadien-3-one	0.2028	1,4-gonadien-3-one	0.2097
	5a-gonan-3-one	0.2031	4-gonen-3-one	0.2122
	5a-gonan-17-one	0.2211	5a-gonan-3-one	0.2221
2-deoxy-*β*-D-ribofuranose	*β*-D-ribofuranose	0.2292	*β*-D-glucopyranose	0.2223
	β-D-glucopyranose	0.2368	α-D-fructofuranose	0.2317
	α-D-fructofuranose	0.2480	α-D-mannopyranose	0.2437
	α-D-galactopyranose	0.2616	*β*-D-ribofuranose	0.2445
	α-D-mannopyranose	0.2696	α-D-glucopyranose	0.2575
lysine	arginine	0.6615	arginine	0.6617
	ethanolamine	0.7882	ethanolamine	0.7621
	choline	1.2682	choline	1.2442
	D-Threose	1.5332	D-Threose	1.4601
	D-xylose	1.5667	D-xylose	1.4912
adenine	pteridine	0.4025	benzothiazole	0.3493
	benzothiazole	0.4321	pteridine	0.3816
	guanine	0.4394	thiazole	0.3981
	7*H*-purine	0.4427	7*H*-purine	0.4254
	indene	0.4614	guanine	0.4265

distances of less than 0.1 probability unit, while useful 2D hits are within 0.2 of the query; here, few of the hits are closer than 0.2 probability units to the query, and some have a distance of 1.5 or greater (2.0 is the theoretical maximum distance under this metric). We interpret this to mean that the Tripos fragment database is too small to warrant the application of 2D-MEP searching, and we turn our attention to the comparisons between the fragment database and the much larger and more diverse NCI compound library (discussed in the next section).

The results found when comparing the fragment databases prepared using 250000 reflections per compound were essentially identical to those discussed above, for all combinations of search type (1D or 2D) and metric (L_1 or R_1) considered (data not shown). This indicates that 50000 reflections per compound ensures adequate convergence, at least for molecules found in the Tripos fragment database.

CPU times for the Tripos fragment database self-comparison (235 queries, 55225 comparisons) under the L_1 metric on a 1.5 GHz Pentium processor were 20.5 s (1D search) and 53.7 s (2D-MEP search). Timings under the R_1 metric were essentially identical. These total times correspond to approximately 370 μs for a single 1D comparison, 970 μs for a 2D-MEP comparison.

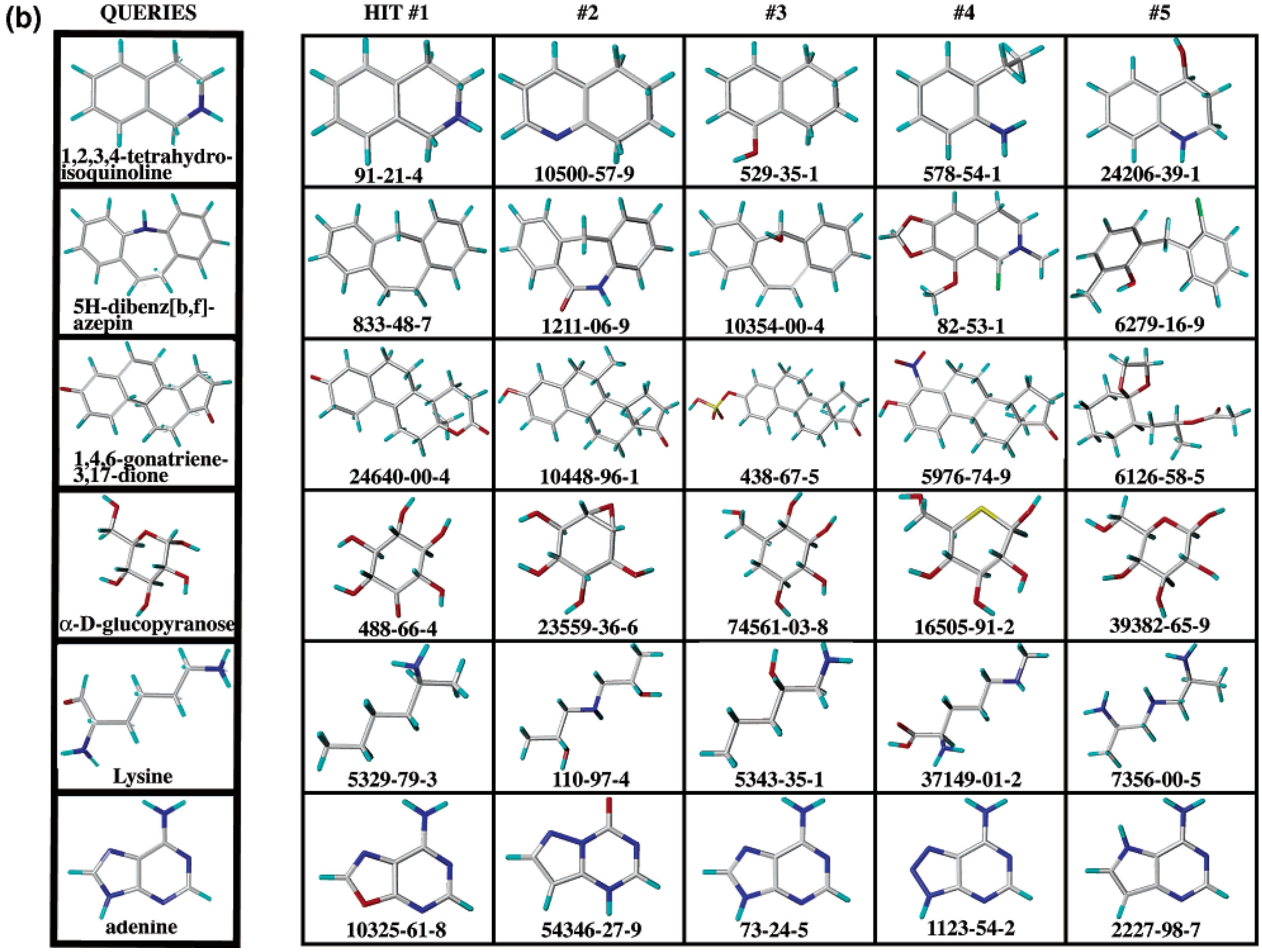
Tripos Fragment Database vs NCI. Table 3a lists top 1D hits for molecules from the Tripos fragment database used as queries against the NCI chemical library. The format of the table follows that of Table 1a, using the same queries and displaying results for 1D shape signatures, but in this table the score columns correspond to the use of different metrics L_1 and R_1, rather than having segment culling enabled or disabled. (Segment culling was used exclusively in preparation of the NCI shape signatures database, so there is no “nonculled” case to compare to.) The hits are labeled by NCI compound IDs (CAS). The top five hit structures for each query are displayed in Table 3(b).

We note at the outset that of the six queries shown, only 1,2,3,4-tetradihydroisoquinoline and adenine are in the NCI database subset used in these searches. In the case of 1,2,3,4-tetradihydroisoquinoline, the NCI entry (CAS #91−21−4) corresponding to the query is selected as the top hit, while for adenine the corresponding hit (CAS #73−24−5) appears in the third position in the hit list. We note in the latter case that the top hit (CAS #10325−61−8) differs from adenine only in the substitution of an amine nitrogen with an oxygen, leading to structures that are very similar in shape; furthermore the scores of #73−24−5 and #10325−61−8 differ by only 0.004 probability units. Given the probabilistic nature of the method and the presence of competing structures of almost identical shape, we feel that it is inevitable that the “best “ chemical structure will not always top the hit list. Also, small differences in the conformation of query and target compounds may influence the order of hits.

The other queries locate target compounds that are generally of very similar structure. Some interesting and initially unexpected hits: for query 1,2,3,4-tetradihydroisoquinoline, the hit #578−54−1 is seen to have the same structure as the query, but with the amine-containing ring opened; for query 5*H*-dibenz[*b*,*f*]azepin, hit #6279−16−9 is similar to the query, but with the central ring opened. This ability to locate “approximate” matches is an interesting feature of the shape signatures approach. We also point out hit #82−53−1 for

Table 3. (a) Results for Six Query Compounds: 1D Shape Signature Comparison of Tripos Fragment Database against the NCI Database using L_1 and R_1 Metrics

	L_1 metric		R_1 metric	
query	hit	score	hit	score
1,2,3,4-tetrahydroisoquinoline	91−21−4	0.0291	91−21−4	0.1153
	10500−57−9	0.0336	10500−57−9	0.1409
	529−35−1	0.0348	578−54−1	0.1428
	578−54−1	0.0380	493−05−0	0.1534
	24206−39−1	0.0397	529−35−1	0.1743
5*H*-dibenz[*b*,*f*]azepin	833−48−7	0.0324	833−48−7	0.1404
	1211−06−9	0.0360	1211−06−9	0.1673
	10354−00−4	0.0415	10354−00−4	0.1789
	82−53−1	0.0441	42263−75−2	0.2142
	6279−16−9	0.0488	51087−02−6	0.2300
1,4,6-gonatriene-3,17-dione	24640−00−4	0.0450	6126−58−5	0.2289
	10448−96−1	0.0556	24640−00−4	0.2561
	438−67−5	0.0570	6968−06−5	0.2672
	5976−74−9	0.0576	20919−82−8	0.2908
	6126−58−5	0.0584	3601−97−6	0.2963
α-D-glucopyranose	488−66−4	0.0546	74561−03−8	0.2223
	23559−36−6	0.0548	488−66−4	0.2548
	74561−03−8	0.0553	488−64−2	0.2548
	16505−91−2	0.0607	6623−68−3	0.2548
	39392−65−9	0.0655	2037−48−1	0.2549
Lysine	5329−79−3	0.0478	37149−01−2	0.1874
	110−97−4	0.0486	6963−39−9	0.1882
	5343−35−1	0.0552	110−97−4	0.2107
	37149−01−2	0.0555	6281−43−2	0.2201
	7356−00−5	0.0563	104−50−7	0.2224
adenine	10325−61−8	0.0271	10325−61−8	0.0944
	54346−27−9	0.0304	54346−27−9	0.0988
	73−24−5	0.0310	5426−35−7	0.1178
	1123−54−2	0.0343	73−24−5	0.1178
	2227−98−7	0.0353	19165−47−0	0.1178

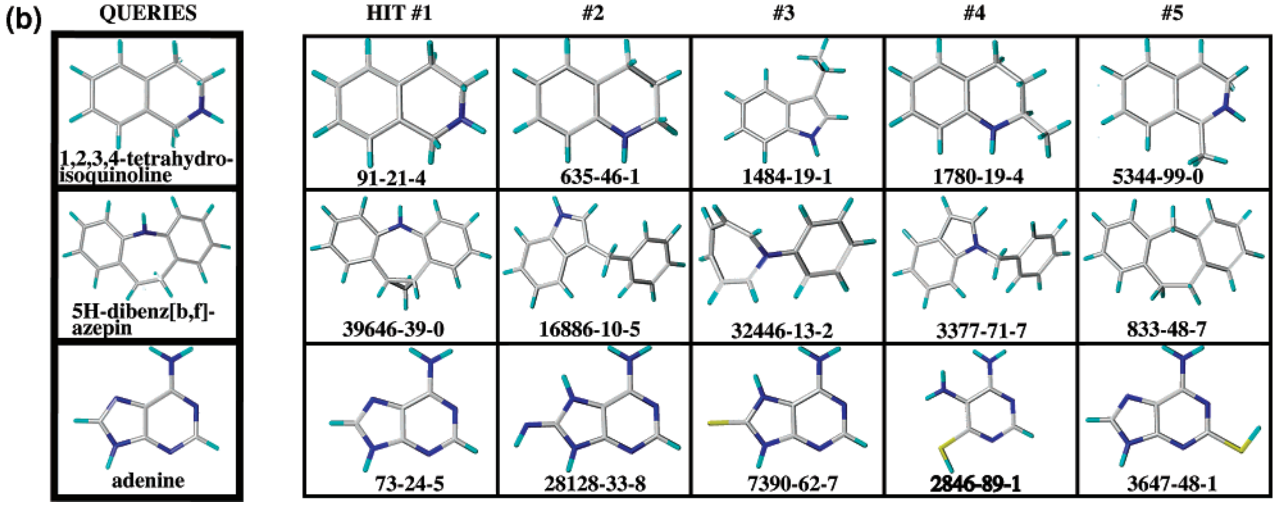


query query 5*H*-dibenz[*b*,*f*]azepin, which clearly bears a weak resemblance to the query despite a low score.

We stress that in any method that “collapses” a large space of chemical structures onto a comparatively small

Table 4. (a) Results for Six Query Compounds, 2D-MEP Shape Signature Comparison of Tripos Fragment Database against the NCI Database using L_1 and R_1 Metric**s**

	L_1 metric		R_1 metric	
query	hit	score	hit	score
1,2,3,4-tetrahydroisoquinoline	91−21−4	0.0701	91−21−4	0.5232
	635−46−1	0.0816	635−46−1	0.6553
	1484−19−1	0.0940	1484−19−1	0.6977
	1780−19−4	0.0983	5344−99−0	0.7295
	5344−99−0	0.1011	1780−19−4	0.8070
5*H*-dibenz[*b*,*f*]azepin	30646−39−0	0.0947	30646−39−0	0.8078
	16886−10−5	0.1079	3377−71−7	0.9075
	32446−13−2	0.1089	16886−10−5	0.9104
	3377−71−7	0.1126	32446−13−2	0.9166
	833−48−7	0.1167	833−48−7	0.9411
1,4,6-gonatriene-3,17-dione	56763−86−1	0.1524	20056−05−7	1.3418
	734−32−7	0.1645	56763−86−1	1.3451
	93998−31−3	0.1682	74924−17−7	1.4169
	20056−05−7	0.1693	734−32−7	1.4949
	74924−17−7	0.1702	71837−43−9	1.5131
α-D-glucopyranose	52019−14−4	0.1815	52019−14−4	1.4065
	49871−87−6	0.1833	58691−27−3	1.4270
	58691−27−3	0.1912	49871−87−6	1.4514
	7404−25−3	0.2015	2280−44−6	1.5418
	14215−77−1	0.2018	14215−77−1	1.5520
Lysine	42021−74−9	0.5473	85385−47−3	4.1381
	58048−33−2	0.5549	58048−33−2	4.2359
	58048−35−4	0.5684	42021−74−9	4.2441
	37082−52−3	0.5719	78582−26−0	4.3301
	78582−26−0	0.5721	62194−88−1	4.3458
adenine	73−24−5	0.0683	73−24−5	0.5048
	28128−33−8	0.1537	28128−33−8	1.0824
	7390−62−7	0.1581	7390−62−7	1.2106
	2846−89−1	0.1744	2846−89−1	1.2491
	3647−48−1	0.1820	1904−98−9	1.2947



descriptor space there *must* be a significant number of false positives, as typified by this example.

Hits under the R_1 metric are similar to those under L_1, involving for some queries the introduction of new hits, but more often simply the reordering in ranking of existing hits. We will not present hit structures for the R_1 metric, since this would largely reproduce the L_1 results.

In Table 4a we present the results for the six query compounds when used in 2D-MEP searches against the NCI database. Despite the much larger size of the NCI database compared to the Tripos fragment database, only three of the queries (1,2,3,4-tetrahydroisoquinoline, 5*H*-dibenz[*b*,*f*]azepin, and adenine) have clearly significant hits (with top scores less than 0.1) and one other (1,4,6-gonatriene-3,17-dione) has hits of borderline significance (with scores all larger than 0.15). Structures for the three low-scoring queries are shown in Table 4b.

Examination of these hits and comparison to Table 3b illustrate the role that electrostatics plays in selecting compounds under a 2D-MEP search. The hits for 1,2,3,4-tetrahydroisoquinoline all exhibit an electronegative nitrogen in a position identical or close to that found in the query. This is not the case for the 1D search (Table 3b, first row), where there is less consistency in the appearance of electronegative atoms in the hits. (We point out in both the 1D- and 2D-MEP searches, the query compound appears as the top hit.) For 5*H*-dibenz-[*b*,*f*]azepin, the top 2D-MEP hit has somewhat weaker shape similarity compared to the top 1D hit, including a cyclopropyl motif not found in the query; at the same time it includes an electronegative nitrogen at a position homologous to that of the query. The top hit is followed by hits that arguably exhibit weaker shape similarity to the query than some of the corresponding 1D hits, but which include an electronegative nitrogen at a position similar to the query. Finally, hit #5 for the 2D-MEP search is the compound found as hit #1 in the 1D search. For adenine, the 2D-MEP search produces hits which in every case contain nitrogens at positions homologous to the query. Compound #2846−89−1 is an interesting "substructure match". The top hit in this case is the query, while in the 1D search the query molecule appears as the #3 hit.

Searching our NCI shape signature database (113,331 compounds) using a 1D query required on average 133 s on a 450 MHz Pentium-III processor running the Linux operating system. This corresponds to 1.17 ms per comparison (which should be compared to the figure of 370 μs quoted above for a 1.5 GHz machine). The average time per comparison for a 2D-MEP search was 3.7 ms.

Score Distributions. A special concern is the distribution of scores for 1D- and 2D-MEP searches against a large and diverse database such as NCI. We have already indicated the range of scores that appear to indicate close similarity between query and target under the L_1 metric; for 1D searches, a distance of 0.05 or less usually corresponds to strong shape similarity, while the range 0.05−0.1 is a borderline region where interesting hits may be mixed with "substructure" matches. For 2D-MEP searches under L_1 the corresponding ranges are 0−0.1 and 0.1−0.2. We expect that even with a large database such as NCI, there should be a relatively small number of close hits for a given query. In contrast, we expect that the vast majority of target compounds should be unambiguously assigned as weak matches.

We computed score distributions under the L_1 metric for the compounds of Table 3 when used as queries against the NCI database. In Figure 7a we show the distribution of 1D scores for 1,2,3,4-tetrahydroisoquinolineand and in 7b the distributions of 2D-MEP scores for this molecule. These plots include a logarithmic axis for the number of compounds observed for a given range of scores. Figures 8a and 8b show the same distributions but with linear vertical axes. Figure 7 highlights the numbers of compounds observed at the extreme left of the distribution, where structurally interesting matches are expected, while Figure 8 provides a better sense of the shapes of the distributions, which appear to be locally Gaussian but with a significant shoulder. The distributions for the other query molecules are overall very similar to the ones shown.

Given these distributions, we can directly compute the cutoff score needed to select a specified percentage of compounds in the NCI database. For a database of this size, two useful cutoffs are those needed to select 0.1% and 0.01% of the compounds, corresponding to approximately 100 and 10 compounds. Analysis of the distributions for the selected compounds reveals a range of 1D scores of 0.04−0.06 to select the top 0.01 percentile of hits, and a range of 0.06−0.09 to select the top 0.1 percentile. For the 2D-MEP distributions, the ranges are 0.11−0.21 and 0.16−0.33. (The 2D-MEP ranges excludes the values for Lysine, which are 0.59 and 0.69 for the 0.1 and 0.01 percentile cutoffs, respectively. These values are anomalous and merely reflect the fact that the query was positively charged while the target database contains mostly neutral compounds.) The cutoff ranges we observe are in reasonable agreement with the qualitative values based on the self-comparison of the Tripos fragment database.

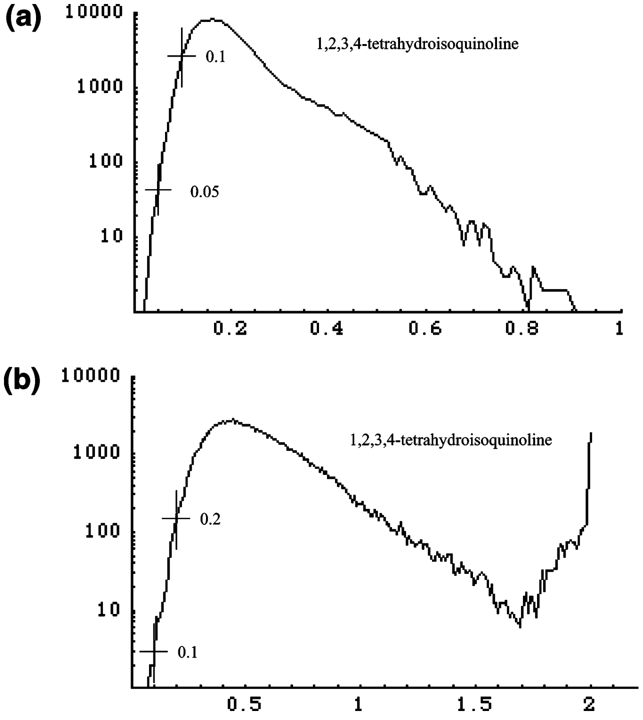


Figure 7. Score Distributions for (a) 1D and (b) 2D-MEP searches against the NCI Database. Vertical axis is number of observed hits; horizontal axis is score. (Logarithmic vertical axis.)

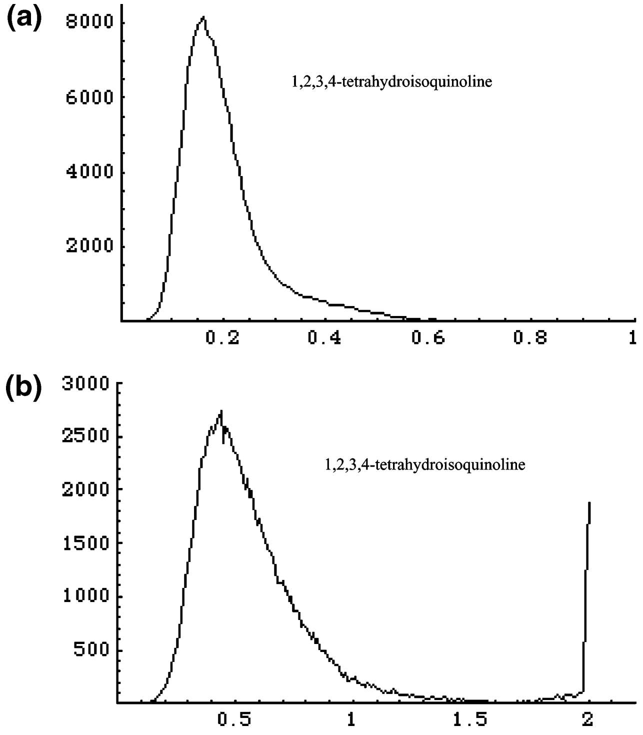


Figure 8. Score Distributions for (a) 1D and (b) 2D-MEP searches against the NCI Database. Vertical axis is number of observed hits; horizontal axis is score. (Linear vertical axis.)

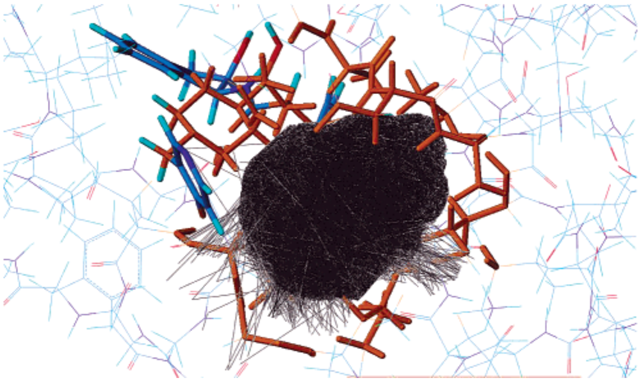


Figure 9. Ray-traces in HIV protease subsite R_2 (as defined in the text). Protein atoms involved in defining a site are orange; framework atoms are colored by atom type. All subsite atoms appear in capped-stick rendering.

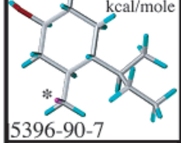
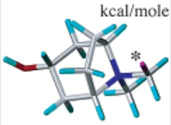
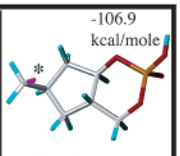
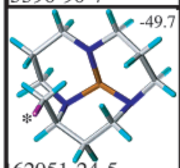
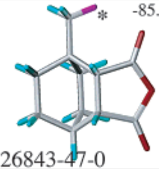
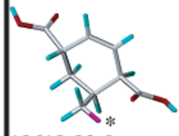
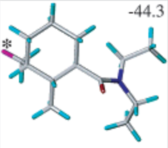
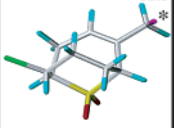
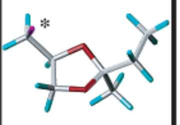
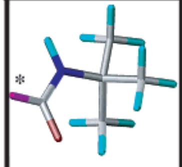
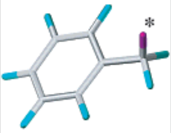
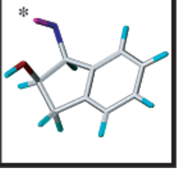
We will further explore the characteristics of the score distributions in later work. For now we point out that the distributions (Figure 8) give the strong impression of being well-represented by a sum of two Gaussians, an observation that may be of practical significance for developing rapid tests for the significance of hits. We also note the secondary peak for the 2D-MEP distribution at the theoretical maximum score of 2.0. We believe this arises from a near-total separation between the distributions of those positively and negatively charged molecules in the database. We will attempt to verify this hypothesis in future work.

Receptor-Based Design. Figure 9 shows the receptor subsite created by removing the R_2 substituent (*tert*-butyl formamide) from the Indinavir framework, along with the associated ray-trace. As described above, subsites were created using the same approach at the R_3 and R_4 positions (not shown).

The top fifty hits for each of the subsite queries were examined, and it was found that in each hit set there were many examples of closely related structures (this was especially the case with the R_4 list). Subsets of structures with high similarity were identified in each list, and only one representative compound from each set was retained in an effort to assemble a structurally diverse group of NCI hit compounds for each subsite. This yielded 27 compounds for R_2, 40 for R_3 and 12 for R_4. After exploding each hit by assignment of all possible attachment points, there were a total of 377 fragments for R_2, 275 for R_3 and 108 for R_4. Each fragment was attached to its target inhibitor site and optimized as described above. Our selection of NCI hit compounds implied a total of 11196900 possible inhibitor structures. Table 5 shows the best three fragments for each subsite, ranked by FlexiDock energy realized after attachment and optimization, along with the CAS ID number for the source NCI compound. We also show for comparison the substituent found at the same position in Indinavir.

To construct a collection of trial inhibitor structure, the best 10 fragments for each of the three variable sites were selected, and inhibitors were constructed using all possible combinations of the selected fragments. Each selected fragment was attached to its target site with the FlexiDock-optimized conformation determined in the first phase of the procedure. This produced 1000 initial structures. Interaction- and self-energies of all the compounds were computed by a utility in ALMS, which continued by ranking the compounds in order of ascending energy. The best fifty compounds were selected for energy minimization in the field of the frozen receptor.

Table 5. Best NCI-Derived Fragments for Sites R1, R2, and R3[a]

	R2	R3	R4
#1	-68.7 kcal/mole  5396-90-7	--94.2 kcal/mole  18650-61-8	-106.9 kcal/mole  73581-87-0
#2	-49.7  62051-24-5	-85.6  26843-47-0	-75.6  19618-93-0
#3	-44.3  54243-68-4	-82.3  15298-66-5	-72.6  2916-28-1
Indinavir			

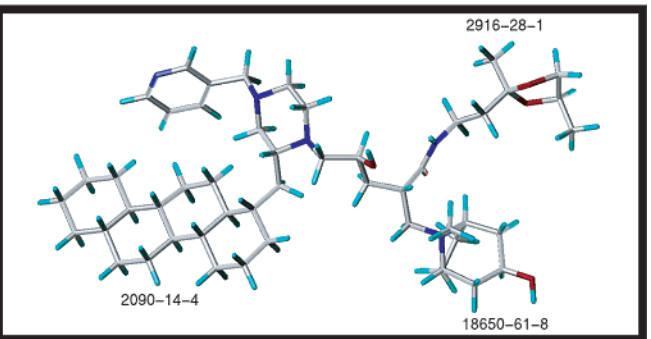
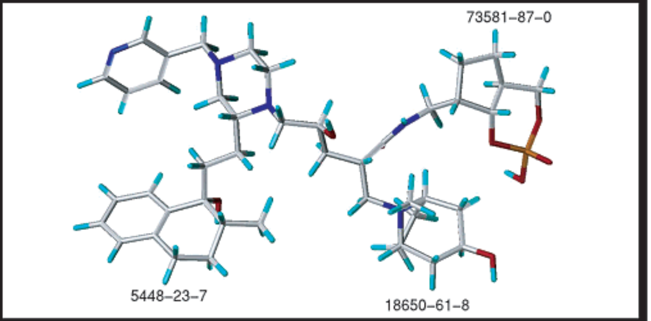
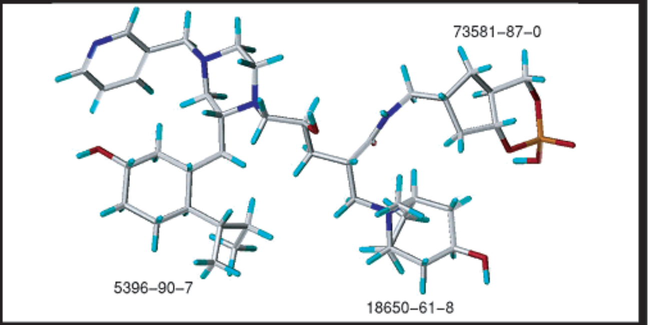
[a] CAS number of source compound lower-left; optimized Energy (kcal/mol) upper-right. Corresponding groups in Indinavir shown for comparison. Attachment points indicated by "*".

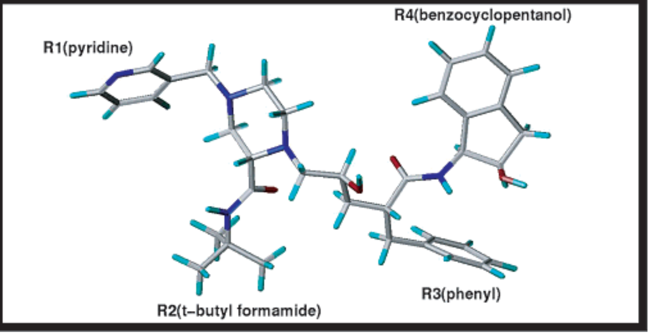
The last phase of this "semiautomated" inhibitor design included generating a rough estimate of binding energy for each of the best fifty inhibitors. This involved removing the inhibitor (with optimized geometry) from the receptor and allowing it to minimize in isolation. Subtracting this minimized energy from the self-energy of the compound when docked provided an estimate of inhibitor strain energy, and this positive quantity was then added to the optimized inhibitor−receptor interaction energy to provide a binding energy estimate. Obviously this simple estimate did not take entropic factors into account nor receptor flexibility.

All computations with SYBYL were carried out on a Silicon Graphics Fuel workstation. Total computing time can be divided into that required for the following phases: Attachment of 760 fragments to their respective inhibitor sites, followed by optimization using FlexiDock, 5.45 h; generation of 1000 trial inhibitors, and initial computation of interaction energy, 1.49 h; final minimization (1000 steps max.) of best 50 inhibitors, 4.7 h. Total CPU time was thus approximately 11.6 h. The time required to scan the NCI database using the receptor-based shape signature queries was approximately 9 min (carried out on 550 MHz Pentium-III processors running Linux).

Three representative inhibitors proposed by this procedure are shown in Table 6, along with their estimated binding energies. Indinavir is included for comparison (it's binding energy estimate was derived from the crystal structure, using the same protocol

Table 6. Selection of Best Inhibitors Constructed using Shape Signatures and ALMS[a]

Rank	Energy(kcal/mol)	Structure
#1	-117.3	
#2	-117.0	
#4	-115.2	

Indinavir	-97.2	 R4(benzocyclopentanol)
		R1(pyridine)
		R3(phenyl)
		R2(t–butyl formamide)

[a] Energies are after 1000 steps (max.) maximization with MaxMin2 and corresponding to a binding-energy estimate as described in the text.

described above). Most of the top-ranking inhibitors involve combinations of a small selection of substituents at the three variable sites; all of the best-scoring compound had the same fragment (derived from compound #18650−61−8) at the R_3 position. Twenty-three of the inhibitors designed by our procedure have an estimated binding energy lower than −100 kcal/mol and are predicted to be better binders than Indinavir.

Discussion

Since shape signatures is a new method, we have presented here a fairly detailed description of the effects of various assumptions that may be employed, such as the inclusion or exclusion of the segment culling procedure, number of reflections generated when preparing the signatures, the incorporation or omission of electrostatic information, and the choice of a metric to be used when making comparisons. While many of these results would have been left out in the discussion of a more mature technique, here they play a valuable role in demonstrating the sensitivity of the method to the choices made for various parameters.

First, it is clear from the results presented here that the method works well in selecting compounds on the basis of shape. (We point out that another laboratory has adopted and extended our approach, also with good results.[35,36]) In the case of the Tripos fragment database self-comparison, the method not only produces a compound of the same or similar chemical class as best hit for a given query, it is generally observed that all compounds of the class appear at or near the top of the hit list (Table 1). This is true under both metrics considered, with hit rankings largely unchanged upon replacing L_1 scores with R_1 scores, which are uniformly larger in magnitude. Furthermore, the results are affected only slightly by inclusion of the segment culling procedure. This gives the strong sense that the method is robust, with results that are relatively insensitive to the detailed choice of parameters. It also calls into question the utility of segment culling, which appears to have only a small impact on the results, in most cases merely changing the rank order of hits. Segment culling is an attractive (albeit computationally expensive) idea and was introduced early in the development of the method; it was used in the generation of the shape signature augmentation to the NCI database used here. However, comparison with results generated with this feature turned off indicate that it may represent a waste of computing resources. (We note that the distributions generated with and without culling are significantly different and cannot be “mixed and matched”; thus, once the NCI database was generated using this feature, query databases destined to be compared against NCI also needed to be generated with segment culling enabled.) Similarly, the insensitivity of results to the choice of metric makes it likely that future work will focus on the simpler L_1 form.

Comparison against the NCI database using the Tripos compounds as queries likewise produced a set of close structural matches for each of the queries, with shape similarity clearly correlated to a small 1D score. Noteworthy are the appearance of approximate and substructure matches as the scores increase; this is a feature of shape signatures comparisons that we have noted in other contexts and will comment on at greater length elsewhere. A general impression is that as scores increase, one observes first close matches, then hits that correspond to substructures or rearrangements of the query, and finally to hits that exhibit no clear similarity to the query. We feel that this is a positive feature of the method, since it is often desirable to identify compounds that have at least partial similarity to an active compound and which may be able to mimic a subset of the interactions of the query with a target receptor. Repeating our experience with the Tripos database self-comparison, the choice of metric had little impact on the results; the only exceptions were observed in those cases (e.g. lysine under 2D-MEP searching) where there were no strong matches to start with.

The inclusion of electrostatics is shown to have significant impact on the hits collected for a given query. In the case of the Tripos fragment database self-comparison, there is inadequate chemical diversity for electrostatics to make a meaningful difference in the search, but when comparing against the NCI database, the 2D-MEP signatures lead to the selection of compounds that have significantly greater electrostatic similarity to the queries. This is reflected in the appearance of atoms with high partial charge at positions in the hits similar to the query. It is also leads to a much smaller number of meaningful hits. This is a consequence of shape and electrostatic information being given equal emphasis in the scoring, leading to a very stringent criterion for the identification of a close match. For example, lysine, a positively charged molecule, finds no close 2D-MEP matches in our version of the NCI database, after finding many shape-similar compounds under a 1D search. On the other hand, using adenine as a query locates compounds that are similar in both shape and polarity under a 2D-MEP search. Moreover, compounds with electrostatic features similar to the query are promoted in the hit list over molecules that exhibit shape-similarity only. In future work, we will explore modified procedures that allow adjustment of the emphasis given to shape and electrostatics when carrying out searches. One useful way to achieve this might be to first screen the database compounds against a query on the basis of shape alone and then to apply 2D-MEP comparisons to reorder the hits based on electrostatic similarity.

It is obviously important to have some criteria for assessing the significance of the hit scores produced in a shape signatures search. The scores for 1D- and 2D-MEP searches against the NCI database are not normally distributed, and it is inappropriate to use z-scores to test for the significance of hits. However, for the queries that we have considered here, meaningful hits appear in the extreme tail of the distribution, and the typical score cutoffs we have proposed lead to selection of a very small percentage of compounds from the database. We stress that while the distributions of scores are of intrinsic interest, as a matter of practice the user of shape signatures decides at the outset how many hits to retain in a given search. From this viewpoint, evidence as to the range of scores likely to correspond to close matches is helpful in determining if the number of hits collected was appropriate. The most important observation concerning the distributions is that close matches between compounds, under either 1D or 2D-MEP searches, are found far from the median, and that it appears to be possible to apply score cutoffs in a reasonably consistent manner to select interesting hits.

Turning to the receptor-based application presented here, we note that shape signatures was used only to collect the raw material to use with ALMS and could just as well have been used to “feed” any of a number of other design strategies. Most of the computational effort was expended on fragment reorientation and energy minimization. Nonetheless the contribution of shape signatures in selecting fragments complementary in shape to the receptor subsites we defined was clearly a critical step. We point out that we could have simply used the entire active site as query, and we made initial attempts at this; however, it soon became apparent that only the interior of the binding site presented a well-defined target for the ray-tracing procedure, so that some approach would need to be developed to “cap” the ends of the active site channel. Including the Indinavir framework in the site thus provided a means to delineate smaller, well-defined regions for ray-tracing and shape signature computation. Furthermore, the relatively small size of our compound database (∼113000 molecules) suggested a greater chance of success in matching smaller subsites, as opposed to finding a close fit to the entire channel. Of course, modifying an existing active compound is a popular method for generating new leads, and our approach falls under this well-respected tradition.

While our receptor-based strategy produced a number of interesting compounds, there are some important caveats with this procedure. First, it is clearly less straightforward to develop a receptor-based approach, since active sites differ dramatically in shape, and a method for restricting the ray-trace to a region of interest may not always be immediately apparent. While our approach of using a framework to define subsites should be widely applicable, even in the case of our demonstration calculation it was necessary to omit one site (R_1) from consideration due to its large solvent accessibility. Second, our approach does not take into account synthetic feasibility; fragments are attached to the framework with no regard to the existence of a synthetic route for preparing the derivative and with no regard to the cost or availability of the necessary reagents. Third, we are currently limited to using shape signatures in 1D mode to select fragments, since shape similarity can be used directly to select compounds complementary in *shape* to a receptor site, but *not* complementary in electrostatic potential, at least assuming our current metrics. All of these issues will be addressed in future work. (We stress that while electrostatics was not taken into account in selection of fragments by shape signatures, it *was* taken into account by ALMS when the fragments were finally attached and their orientations optimized.)

Perhaps the most important unanswered question concerning shape signatures is the influence of conformation on the identification of similar compounds, an issue which we have not addressed in this report. While we have some evidence that comparisons made using shape signatures are not extremely sensitive to conformational differences, we nonetheless recognize this as an important concern and will discuss this question in detail elsewhere. We would point out that given the speed of the method, this issue is relatively easy to address. For a given query, one can easily generate alternate conformers and use these separately to scan a target database. Provided that one of the query conformers is found in the database, a close match is sure to be identified.

We finally note that even in the case of our receptor-based strategy, shape signatures is a comparatively easy technique to apply. Especially for ligand-based applications the method does not require extensive experience in constructing queries, adjustment of numerous parameters, or sophistication in the interpretation of results, which can be serious drawbacks with other methods. Moreover, it directly addresses those features of molecules, namely their shape and surface properties, which are most critical to determining their biological activity. In this it is more efficient than those methods that focus on chemical structure, where shape and electrostatics are merely implied by chemical connectivity, and where one must take care to identify compounds that, while different in chemical structure, are closely similar in shape.

Conclusions

We have presented a new method, shape signatures, which we feel has great promise in the area of computer-aided molecular design. The technique focuses on shape rather than chemical structure, is independent of molecular orientation, and permits very rapid screening of large databases for compounds with properties similar to a query molecule. We have demonstrated that the method works well in selecting molecules on the basis of shape and polarity and have applied it in a receptor-based strategy where it led to the construction of compounds predicted to have better binding energy than the initial lead.

While the method is useful as it stands, we are working to address a number of issues that need to be resolved to extend the range of application of the technique. Chief among these are the incorporation of conformational flexibility into the method and for application in receptor-based strategies the development of a new metric that can handle electrostatic complementarity. We will also seek to introduce more "chemical intelligence" into our approaches, especially as regards identifying fragments that will be compatible with known synthetic strategies.

References

(1) Klebe, G. Recent developments in structure-based drug design. *J. Mol. Med.* **2000**, *78*, 269−281.

(2) Waszkowycz, B. Structure-based approaches to drug design and virtual screening. *Curr. Opin. Drug Discovery Dev.* **2002**, *5*, 407−413.

(3) Xue, L.; Bajorath, J. Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. *Comb. Chem. High Throughput Screen.* **2000**, *3*, 363−372.

(4) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.

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

(6) Raymond, J. W.; Willett, P. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 521−533.

(7) Thorner, D. A.; Willett, P.; Wright, P. M.; Taylor, R. Similarity searching in files of three-dimensional chemical structures: representation and searching of molecular electrostatic potentials using field-graphs. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 163−174.

(8) Polanski, J.; Walczak, B. The comparative molecular surface analysis (COMSA): a novel tool for molecular design. *Comput. Chem.* **2000**, *24*, 615−625.

(9) Hull, R. D.; Singh, S. B.; Nachbar, R. B.; Sheridan, R. P.; Kearsley, S. K.; Fluder, E. M. Latent semantic structure indexing (LaSSI) for defining chemical similarity. *J. Med. Chem.* **2001**, *44*, 1177−1184.

(10) 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.

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

(12) Putta, S.; Lemmen, C.; Beroza, P.; Greene, J. A novel shape-feature based approach to virtual library screening. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1230−1240.

(13) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Recent advances in comparative molecular field analysis (CoMFA). *Prog. Clin. Biol. Res.* **1989**, *291*, 161−165.

(14) Cramer, R. D. Topomer CoMFA: a design methodology for rapid lead optimization. *J. Med. Chem.* **2003**, *46*, 374−388.

(15) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **1982**, *161*, 269−288.

(16) Ewing, T. J.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 411−428.

(17) Miller, M. D.; Kearsley, S. K.; Underwood, D. J.; Sheridan, R. P.FLOG: a system to select 'quasi-flexible' ligands complementary to a receptor of known three-dimensional structure. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 153−174.

(18) Jones, G.; Willett, P.; Glen, R. C. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.* **1995**, *245*, 43−53.

(19) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727−748.

(20) Tripos, Inc., St. Louis, MO.

(21) Richards, F. M. Areas, Volumes, Packing and Protein Structure. *Annu. Rev. Biophys. Bioeng.* **1977**, *6*, 151−176.

(22) Connolly, M. L. Solvent-accessible Surfaces of Proteins and Nucleic Acids. *Science* **1983**, *221*, 709−713.

(23) Connolly, M. L. Analytical Molecular Surface Calculation. *J. Appl. Crystallogr.* **1983**, *16*, 548−558.

(24) Zauhar, R. J. SMART: a solvent-accessible triangulated surface generator for molecular graphics and boundary element applications. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 149−159.

(25) Sitkoff, D.; Sharp, K. A.; Honig, B. Accurate Calculation of Hydration Free Energies Using Macroscopic Solvent Models. *J. Phys. Chem.* **1994**, *98*, 1978−1988.

(26) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity − Rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219−3288.

(27) Milne, G. W. A.; Miller, J. A. The NCI Drug Information System. 1. System Overview. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 154−159.

(28) Milne, G. W. A.; Feldman, A.; Miller, J. A.; Daly, G. P.; Hammel, M. J. The NCI Drug Information System. 2. DIS Pre-Registry. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 159−168.

(29) Milne, G. W. A.; Feldman, A.; Miller, J. A.; Daly, G. P. The NCI Drug Information System. 3. The DIS Chemistry Module. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 168−179.

(30) Milne, G. W. A.; Miller, J. A.; Hoover, J. R. The NCI Drug Information System. 4. Inventory and Shipping Modules. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 179−185.

(31) Zehnacker, M. T.; Brennan, R. H.; Milne, G. W. A.; Miller, J. A. The NCI Drug Information System. 5. DIS Biology Module. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 186−193.

(32) Zehnacker, M. T.; Brennan, R. H.; Milne, G. W. A.; Miller, J. A.; Hammel, M. J. The NCI Drug Information System. 6. System Maintenance. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 193−197.

(33) Moyna, G.; Welsh, W. J.; Zauhar, R. J. Automating Drug Design Against Mutable Targets. Presented at the National Meeting of the American Chemical Society, Anaheim, CA, 1999.

(34) Chen, Z.; Li, Y.; Chen, E.; Hall, D. L.; Darke, P. L.; Culberson, C.; Shafer, J. A.; Kuo, L. C. Crystal structure at 1.9 Å resolution of human immunodeficiency virus (HIV) II protease complexed with L-735, 524, an orally bioavailable inhibitor of the HIV proteases. *J. Biol. Chem.* **1994**, *269*, 26344.

(35) Breneman, C. M.; Sundling, C. M.; Sukumar, N.; Shen, L.-L.; Katt, W. P.; Embrechts, M. J. New developments in PEST shape/property hybrid descriptors. *J. Computer-Aided Mol. Des.* **2003**, *17*, 231−240.

(36) Whitehead C. E.; Sukumar, N.; Breneman, C. M. Transferable Atom Equivalent Multi-Centered Multipole Expansion Method, *J. Comput. Chem.* (Special Issue on electron densities and electrostatic potentials;. Gadre, R. S., Ed.) **2003**, *24*, 512−529.

JM030242K