Chemical Function Queries for 3D Database Search

Jonathan Greene,* Scott Kahn, Hamid Savoj, Peter Sprague, and Steven Teig

Molecular Simulations, Inc., 555 Oakmead Parkway, Sunnyvale, California 94086-4023

Received March 30, 1994®

The advent of three-dimensional (3D) molecular database searching motivates this investigation of how best to formulate queries for compounds likely to bind to an enzyme or receptor. 3D queries in the literature generally refer to simple topological features (e.g., nitrogen or phenyl). To better capture the chemist's intent and to find functionally equivalent but structurally diverse compounds, generalized chemical function definitions are proposed for hydrogen bond acceptors/donors, charge centers, and hydrophobes. Use of these function definitions in 5-HT$_3$ antagonist and ACE inhibition queries is shown to identify dramatically more hits capable of forming the hypothesized interactions. Furthermore, false positives that have features inaccessible to the receptor are eliminated. Next, the literature on intermolecular interaction energy is reviewed to determine what geometric tolerances are chemically reasonable in queries. Finally, it is shown that the commonly used distance constraints poorly distinguish conformers that do and do not superimpose well with receptor features. An alternative, the *location constraint*, is proposed. Queries for angiotensin II antagonism and HLE inhibition are described and used to search 203 000 compounds.

INTRODUCTION

Systems for searching a database of molecules based on their 3D characteristics[1−4] are becoming established as a useful tool in drug discovery.[5] This paper discusses how to formulate search queries for the purpose of identifying molecules likely to exhibit binding activity to some enzyme or receptor.

1. How should the key molecular features be described? Detailed chemical function definitions are proposed that are more general and accurate than the simple topological definitions commonly used today.
2. What geometric tolerances are appropriate? The literature on intermolecular interaction energies is reviewed to provide quidelines for choosing chemically reasonable tolerances.
3. What form of geometric constraint is best? Some deficiencies of constraints on interatomic distances, commonly used in published queries, are investigated. An alternative based on RMS superposition is proposed.

Most 3D search queries in the lierature define molecular features by atomic topologies such as a nitrogen, a phenyl ring, or a carbonyl oxygen.[4,6,7] While simple topologies are convenient for search algorithms, they do not describe all groups that can serve the same chemical function in a binding interaction. A chemist proposing a structure−activity hypothesis can often suggest what chemical function a topological feature is serving, for example, hydrogen bond donor site,[8] lipophilic site,[8] basic center,[8,9] and so on. A query that captures these general chemical functions can identify novel structures that a simple topological query misses.

An early effort in this direction was the ALADDIN system.[1] Concurrent with our work, Bush and Sheridan[10] described a rule-based method for classifying individual atoms into functional types. In section 1 below we describe chemical function definitions for hydrogen bond acceptors/donors, charge centers, and hydrophobes that extend previous work in several ways:

* use of complex logical expressions to capture various cases correctly
* more accurate non-atom-centered representation of π-delocalized charge systems such as carboxyl and guanidinium
* identification of contiguous solvent-accessible patches of hydrophobic surface, including chain as well as ring structures
* If an atom is to participate in a hydrophobic or hydrogen-bonding interaction, it must be exposed on the ligand surface.[11,12] Checking this eliminates many false positives that violate the intent of the query.

Examples are given that quantify the impact of generalized function definitions on search results.

Once features have been determined, geometric constraints must be added to the query and tolerances set. Martin et al.[13] report that pharmacophore identification works best with tolerances of 1−2 Å on distances. On the other hand, many queries used in the literature to benchmark database search systems have tolerances on interatomic distances of ±0.1 Å or less.[4,7] Guidance in choosing physically reasonable geometric tolerances for database queries can be obtained from crystallographic data and from the dependence of interaction energies on the relative positions of the atoms involved. Section 2 gives estimates of reasonable tolerances on these grounds, distinguishing each particular class of interaction.

The enthalpy of ligand binding is in part determined by whether key atoms (or groups) in some conformer of the ligand can superimpose with corresponding ideal positions based on the receptor. The rationale for this is that the energy of each individual interaction is a function of the relative positions of the ligand and receptor atoms involved in the interaction.[14] A simple and commonly used measure of superimposability is the minimized root mean square (RMS) distance between pairs of corresponding points, one

® Abstract published in *Advance ACS Abstracts*, October 15, 1994.

determined from ligand coordinates and one from receptor coordinates, after optimal alignment.

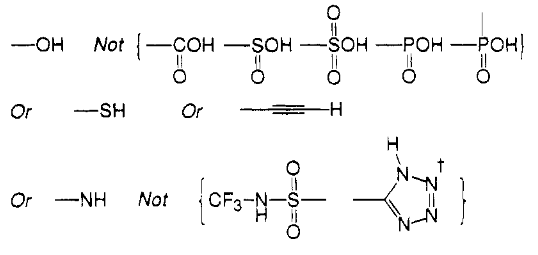
On the other hand, most 3D search systems support only internal constraints, i.e., constraints among points determined exclusively from the ligand coordinates. Perhaps this is because internal constraints are computationally simple.[1] Constraints on the distance between two points (generally atoms) in the ligand are most common, but angles and torsions can also be involved. Clearly, distance constraints alone are insufficient (in the case of chiral molecules) to select conformers that superimpose well with a receptor because a conformer and its mirror image exhibit the same interatomic distances. But even when the query is not enantioselective or the molecules are achiral, internal constraints poorly capture superimposability. The creators of the ALADDIN[1] and 3DSEARCH[2] systems recognized the limitations of internal constraints in the context of placing excluded volumes and provided a means to do this by RMS fit to ideal positions. However, even these systems rely on internal constraints for required features. In section 3 we demonstrate the importance of constraining superimposability rather than distances for required features and examine the cause of the discrepancy between these two approaches. As an alternative to distance constraints, we propose the use of *location constraints*, which directly reflect superimposability. Location constraints also provide a more intuitive way to view and edit queries.

The searches reported in this paper were carried out using Catalyst/Info Release 2.2[15] (fast mode) and the CHM query language.[16] The searches are conducted on the BioByte-MasterFile database (a set of 24 416 compounds of pharmaceutical interest collected by the Pomona Medicinal Chemistry Project), the NCI database (a set of 99 506 compounds collected at the National Cancer Institute), the Derwent World Drug Index (37 970 compounds), and the Maybridge database (41 742 compounds).[17] The method for constructing the databases attempts to cover the low energy conformational space of the molecules by judicious sampling using the poling technique.[18] Up to 120 representative conformers per molecule are used. Where there are chiral centers of unspecified chirality, all possible stereoisomers are considered. Conformers matching the query are required to be within 20 kcal of the estimated global minimum energy using a force field based on CHARMm.[19]

1. CHEMICAL FUNCTION DEFINITIONS

The major forces involved in selective binding (molecular recognition) are hydrogen bonding, electrostatics, and hydrophobic interactions.[12,20,21] In this section we propose generalized functional definitions for groups able to participate in such interactions. Queries can be constructed using these definitions as they are, or, for unusual applications, the definitions can be modified or supplemented via the CHM query language[16] in which they are encoded.

1.1. Hydrogen Bond Acceptors and Donors. We consider any nitrogen, oxygen, or sulfur atom with at least one available (i.e., nondelocalized) lone pair to be an acceptor atom. The lone pair of hypervalent sulfur is not considered available. Hydrogen bond acceptors are further partitioned based upon the environment: for nonlipid environments, we exclude all basic amines, which are prontonated at physiological pH.



† All tautomeric forms

Figure 1. Definition of hydrogen bond donor.

Hydrogen bond donor atoms are identified by the availability of an electropositive hydrogen atom. Hydrogen bond donors, as shown in Figure 1, include

1. all hydroxyls that are not contained within a carbon, sulfur, or phosphorous acid
2. all thiols and acetylenic hydrogens
3. hydrogens attached to nitrogens that are not part of a trifluoromethylsulfonamide or tetrazole moiety (which are assumed to be ionized at physiological pH)

For each acceptor atom in the ligand, we must identify the possible positions for the corresponding donor atom of the receptor that permit a good hydrogen bond but do not collide with other parts of the ligand. The same is true for each donor atom in the ligand and its corresponding acceptor. The position of the hydrogen bonding atom in the receptor can be projected from the ligand atom positions.[1] The ligand atom, its symmetry, and the number of nondelocalized lone pairs determine the bond direction. At present we assume a distance of 3.0 Å independent of the atom types. When the ligand atom is at the end of a rotatable bond, we sample all possible positions on the circle swept out by the projected point as the bond rotates; examples are hydroxyl, thiol, and primary amine groups. If there exist nonrotatable hydrogen positions, we sample all possible hydrogen locations. On a secondary amine for instance, we consider both positions obtained by interchanging the lone pair and hydrogen.

To verify that a projected position is accessible, we place a probe sphere of 1.0 Å radius at the point and determine whether it collides with he van der Waals radius of any atom in the conformer. The 1.0 Å radius was chosen to be smaller than the van der Waals radii of hydrogen bonding atoms to allow for the fact that strong hydrogen bonds are formed even when the geometry is not ideal. The collision check excludes about one-fifth of the possible projected positions, reducing the number of false positives from the search.

The definitions shown do not account for all forms of tautomerism, though they could be extended to do so.

1.2. Charge Centers. A simple way to identify charge centers is to find atoms with a nonzero formal charge, but this would be inaccurate in several cases. First, compounds that would be ionized at physiological pH might be stored in the database in their neutral forms; moieties such as carboxylic acids, aliphatic amines, and guanidino groups should thus be considered as charge centers even though they bear no formal charge.

Delocalized charges present additional complications. In the case of π-delocalized systems, such as guanidinium and carboxylate, the orbital itself is delocalized across the atoms. The centroid of the relevant heteroatoms is thus a reasonable indicator of the position of maximum interaction energy. This

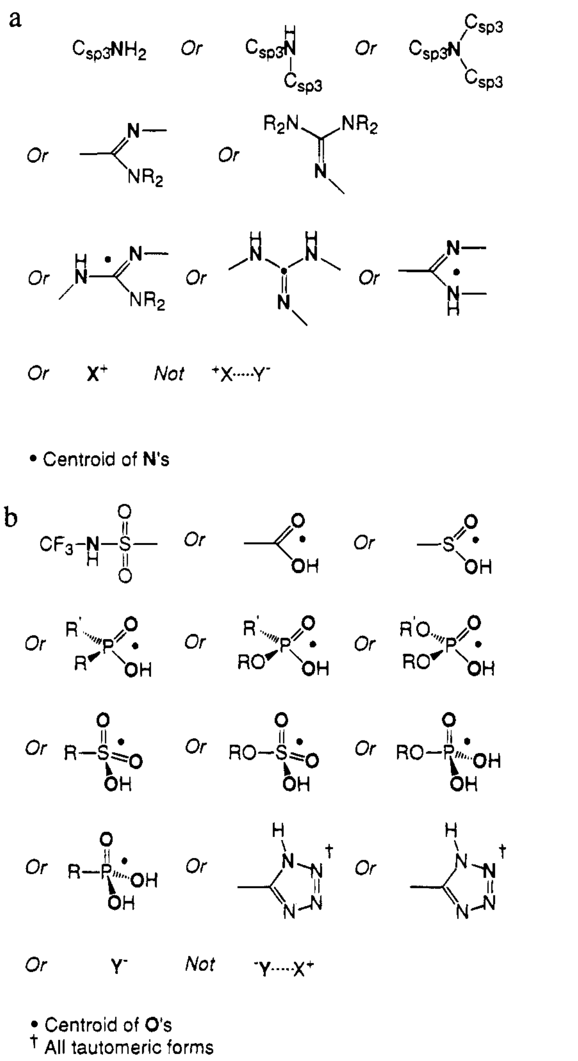


Figure 2. (a) Definition of positive charge center and (b) definition of negative charge center.

is not the case when the charge resides primarily in the σ system. In tetrazole for instance, it is important to localize the charge onto each of the alternative nitrogens individually.

We attempt to accommodate these complications, at least for groups commonly encountered in pharmaceutical applications, with the following definitions. A positive charge center (Figure 2a) includes

1. an atom bearing a formal positive charge if it is not directly adjacent to an atom with a formal negative charge
2. the nitrogen in primary, secondary, and tertiary aliphatic amines
3. the imino nitrogen of *N,N*-disubstituted amidines and *N,N,N,N*-tetrasubstituted guanidines
4. the centroid of the hydrogen-bearing amino nitrogen and the imino nitrogen in *N*-substituted and unsubstituted amidines and *N,N*-disubstituted (substitution on the same amino nitrogen) guanidines with at least one amino hydrogen
5. the centroid of the three nitrogens in guanidines bearing at least one hydrogen on each amino nitrogen

Weaker groups such as imidazoles and pyridines can also be added to the definition if necessary.

A negative charge center (Figure 2b) includes

1. an atom bearing a formal negative charge if it is not directly adjacent to an atom with a formal positive charge
2. the centroid of the oxo and hydroxyl oxygens in carboxylic, sulfinic, and phosphinic acids
3. the centroid of the oxo and hydroxyl oxygens in phosphoric diesters and phosphonic esters
4. the centroid of the two oxo oxygens and the hydroxyl oxygen in sulfuric and sulfonic acids
5. the centroid of the oxo oxygen and the two hydroxyl oxygens in phosphoric monoesters and phosphonic acids
6. the nitrogen in trifluoromethylsulfonamides
7. any nitrogen (i.e., not the centroid) in a non-*N*-substituted tetrazole

Naturally, additional refinements are required to identify charge center groups completely. Examples might include substituted aza aromatics such as pyridines and imidazoles, where electron donor and acceptor groups affect the acidity/basicity of the aromatic nitrogens. A further improvement would be to use partial rather than integral charges. Ultimately of course, one could describe the electrostatic field. Such extensions are left for subsequent reports.

1.3. Hydrophobic Regions. Many queries in the literature identify hydrophobic regions by the centroid of a molecular fragment, often a phenyl or other aromatic ring.[9,22] Our goal is to formulate a more general and accurate definition for a hydrophobic surface region. For instance, methyl, ethyl, and *tert*-butyl groups, short or long carbon chains, and nonaromatic rings might all present hydrophobic surfaces in certain cases. We adopt the common practice, which has at least the virtue of simplicity, of using a single point to capture the position of the hydrophobic group. (Of course, more accurate representations describing the size, shape, and directionality of the group might be advantageous in certain applications. These are being considered for future investigations.)

Previous studies have proposed rules for estimating the contributions made by individual atomic groups to desolvation energy.[23-26] These rules are intended for estimation of partition coefficients and free energies of binding or for use in modeling studies.

Ghose and Crippen[23] assume that the desolvation energy can be expressed as the sum of atomic contributions. Carbon, hydrogen, oxygen, nitrogen, sulfur, and halogens are classified into 110 atom types, and hydrophobic contributions for each type are determined by regression from log *P* data. The plethora of types allows a distinction between atoms that are shielded from solvent by their neighbors and atoms that are not. However, the method cannot account for the possibility that an atom is buried by a topologically distant part of the molecule that folds back.

Eisenberg and McLachlan[24] use a different model that explicitly considers the solvent accessible surface area[27] A_i of an atom. The estimated desolvation energy of a group of atoms is given by

$$\Delta G = \sum_{\text{atoms } i} A_i \Delta\sigma_i$$

where $\Delta\sigma_i$ is estimated from experimental data for five different types of atoms: carbon, neutral oxygen and

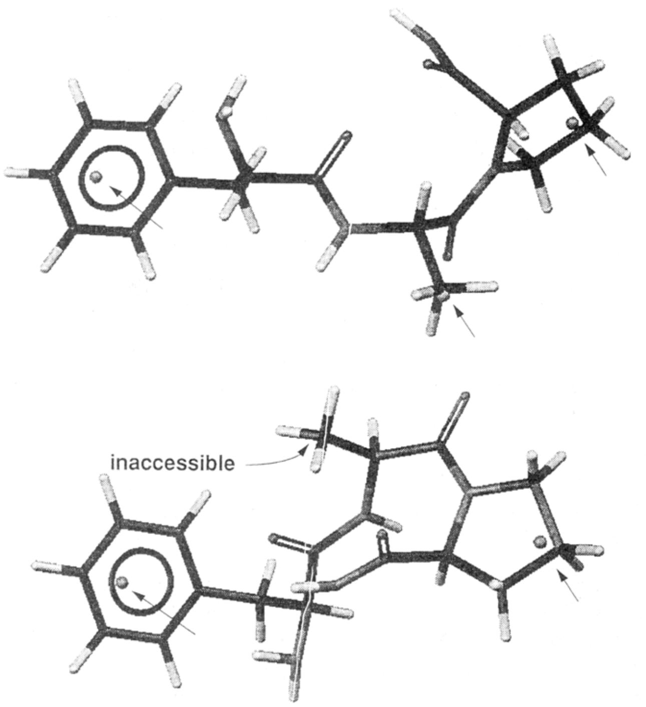


Figure 3. Hydrophobe points (arrows) for two different conformations of Phe-Ala-Pro. The location marked "inaccessible" is a position where the hydrophobic group is partially buried.

nitrogen, negatively charged oxygen, positively charged nitrogen, and sulfur.

Employing either of these approaches in a conventional query language would be quite difficult. A purely topological definition with the complexity of Ghose and Crippen's 110 types would overwhelm most current search systems. On the other hand, Eisenberg and McLachlan's approach requires a mechanism for considering accessible surface area. Even if atomic contributions can be determined, one is still faced with the problem of grouping atoms together into regions. Ideally, the grouping would depend on the surface areas. Furthermore, one can encounter excessive numbers of redundant overlapping groups. For example, a query looking for four adjacent hydrophobic carbons would have $n - 3$ overlapping instances in a carbon chain of length n. A method that produced an adequate sample of these would be more efficient.

Our solution to these problems is to use a construct in the query language specifically for hydrophobes. The method for finding hydrophobes can thus be encoded in the search program itself rather than the query language, allowing the method to be very complex and yet quite efficient. Our algorithm employs Eisenberg and McLachlan's expression above for the desolvation energy of a set of atoms, but we use a more complex system of types that takes into account an atom's neighbors. Once the hydrophobic contribution of each atom has been determined, we group adjacent atoms constituting a significant region of hydrophobic surface. Finally, we mark each region with a point at the surface-weighted centroid of the atoms. The details of the algorithm are in the Appendix.

Figure 3 shows an example of the hydrophobic groups identified in two conformations of Phe-Ala-Pro. In one conformation, three hydrophobic surface regions are identified. In the second, one of them has been partially bured by a phenyl ring and thus is ignored.

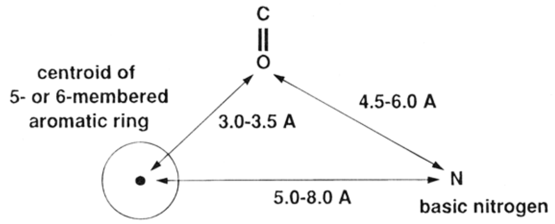


Figure 4. Query A.

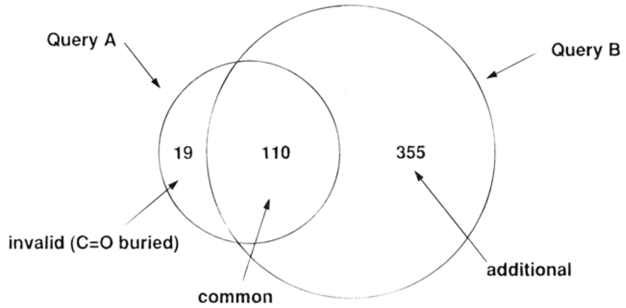


Figure 5. Searching with topological query A (five- or six-membered aromatic ring, C=O, basic N) vs functional query B (five- or six-membered aromatic ring, hydrogen bond acceptor, positive charge center). Query A finds only 110 (24%) of the 465 compounds found by query B.

1.4. Impact of Functional Versus Topological Features.

To illustrate the importance of properly generalizing feature definitions, we consider the 5-HT_3 antagonist pharmacophore proposed by Hibert et al.[9] as interpreted in the MOL query language.[6] This pharmacophore, denoted query A and shown in Figure 4, conists of a five- or six-membered aromatic ring, a carbonyl oxygen, and a basic nitrogen at defined distances. Searching the BioByteMasterFile database with query A, we find 129 compounds.

Hibert et al. suggest that the carbonyl oxygen is serving as a hydrogen bond acceptor and that the basic nitrogen is serving as a charge center. We define query B by substituting the chemical function definitions for acceptor and positive charge center (described above) for the oxygen and nitrogen, respectively, of query A. Searching the same database with query B, we find 465 compounds. As shown in Figure 5, 355 of these were new compounds, found because of the greater generality of the function definitions. On the other hand, 19 of the compounds found by query A failed to satisfy the additional requirement of query B that carbonyl oxygens be exposed on the surface to be considered acceptors. Thus, the simple topological query A misses 76% of the compounds satisfying the generalized functional query B, while finding 15% additional false positives.

Evidence for the importance of a generalized hydrophobe definition can also be found in published pharmacophores. Petrillo and Ondetti[28] suggest that the aliphatic side chain of nLeu in nLeu-Ala-Pro and the aromatic ring of Phe-Ala-Pro, both angiotensin converting enzyme (ACE) inhibitors, interact with the same hydrophobic pocket in the binding site. We consider two versions of a query for ACE inhibition proposed by Sprague[29] and shown in Table 1. Query D consists of a negative charge center, two hydrogen bond acceptors, and three hydrophobes, as defined above. Query C is the same except that hydrophobe 3, corresponding to the nLeu and Phe side chains, is replaced by a more specific topological definition like that used in query A: the centroid of a five- or six-membered aromatic ring.

Table 1. ACE Inhibition Query[a]

function	tolerance	x	y	z
hydrophobe 1	1.6	−2.02	1.26	−0.54
hydrophobe 2	1.6	2.26	1.76	−3.17
hydrophobe 3	1.6	5.09	3.57	2.92
acceptor atom 4	1.6	1.19	−1.20	−1.33
projected point	2.2	3.47	−2.75	−2.51
acceptor atom 5	1.6	5.76	−0.12	−0.80
projected point	2.2	6.61	−1.40	−3.39
negative charge center	1.6	0.66	2.31	1.28

[a] Values in Å.

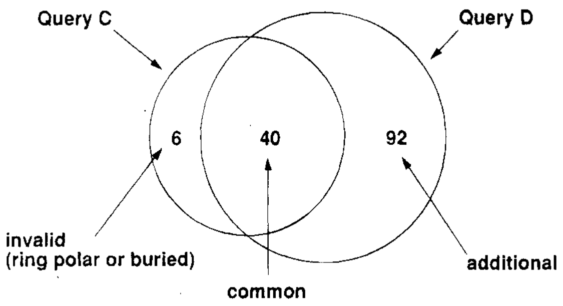


Figure 6. Searching with query C (using five- or six-membered aromatic ring) vs query D (using generalized hydrophobic group). Query C finds only 40 (30%) of the 132 compounds found by query D.

In a search of the BioByte database, query C finds 46 hits. As shown in Figure 6, the fully generalized query D finds 92 additional hits missed by query C. On the other hand, six compounds found by query C are correctly eliminated by query D because the rings identified as hydrophobes are actually polar or not exposed to solvent. Thus query D returns 132 hits overall. Query C misses 70% of the compounds satisfying query D, while finding 15% additional false positives.

We conclude that basing a query on generalized chemical functions instead of atomic topologies greatly enhances the potential for discovering new leads. The generalized function definitions, while still in keeping with the hypothesized binding interactions, yield dramatically more hits. At the same time, compounds with features inaccessible to the receptor are eliminated.

2. APPROPRIATE TOLERANCES FOR GEOMETRIC CONSTRAINTS

The question of what tolerances to use in the geometric constraints of a query is a complex one. When presented with a query having very tight tolerances, say 0.1 Å, it is tempting to imagine that the query is not only precise but accurate in selecting molecules that bind. However, such precision is unjustifiable on physical-chemical grounds. For example, if the energy of an interaction is only weakly affected when the ligand atom moves by half an angstrom, it makes no sense to use a query that constrains that atom to a narrower region.

We consider four important interactions for ligand binding: hydrogen bonding, $\pi-\pi$ interactions, hydrophobic interactions, and charge interactions. For each one we examine the literature regarding the dependence of interaction energy on geometry and try to formulate some conclusions about appropriate tolerances.

Finally, we note that tolerances in queries that are derived from crystal structures of receptors are necessarily limited by the resolution of the structures. This is considered in section 2.5.

2.1. Hydrogen Bonding. A number of studies have investigated the geometric variation of hydrogen bond interactions in condensed phases. Murray-Rust and Glusker[30] found considerable variation in hydrogen bond geometry in their examination of the Cambridge Structural Database (CSD). For an X−H donating group, where X is N or O, they found that the distance between X and the acceptor atom ranges between 2.4 and 3.0 Å. Even greater variations were found in the bond direction. For example, the position of the atom X donating a hydrogen to a carbonyl oxygen was found to vary over an arc of approximately 120° within the plane of the lone pair orbitals and 55° in the perpendicular direction.

Taylor and co-workers independently did a more focused study of amine−carbonyl hydrogen bonds in the CSD.[31] They report a hydrogen-to-acceptor distance of 1.8−2.1 Å. Adding 1.0 Å for a typical N−H covalent bond gives a range of 2.8−3.1 Å. Angle variations of as much as 16 degrees from linearity were common. Other efforts at modeling ligand-receptor binding have used values as high as 3.5 Å for the hydrogen bond distance.[32]

The development by Boobbyer et al.[33] of an empirical hydrogen bonding potential function provides a means of relating geometric tolerances to energy. The distance-dependent component of their potential function consists of a $1/r^6$ attractive and a $1/r^8$ repulsive term. Employing the parameters cited, the potential indicates that a 1 kcal/mol energy threshold allows a range of distances 1.2, 0.9, and 0.6 Å wide for N••H••N, N••H••O, and O••H••O systems, respectively.

Regarding bond direction, Boobbyer et al. give a distribution of the hydrogen bond angle in the lone pair plane for carbonyls. The maxima of this distribution are located at ±60 degrees along the expected lone pairs of the sp^3 hybridized oxygen. There is also an important contribution at 0° from linear hydrogen bond geometries. A 1 kcal/mol energy threshold allows an arc of about 160°. If we apply this energy instead in the perpendicular direction, the arc is about 80° wide.

Similar findings have been obtained in studies of hydrogen bonds in protein systems,[34] so the reported variations are unlikely to be specific to small-molecule crystal structures or artifacts of a specific molecular environment.

To summarize, we can infer the following rough guidelines. Since estimates of hydrogen bond distances range from 2.4 to 3.5 Å, we assume the position of the participating ligand atom can vary by about 0.5 Å from its ideal position with respect to the receptor. We must also account for the angular variation, which might be ±16 to 40 degrees. We can put this in terms of a variation in position by taking the length of the chord defined by the angle at a 3 Å radius. An angular variation of 40 degrees corresponds to a variation in position of 2.1 Å.

2.2. $\pi-\pi$ Interactions. π-Stacking interactions (while somewhat a misnomer) are considered important contributors in interactions involving protein receptors.[35] Preferring an orientation that places the edge of one aromatic system juxtaposed with the π system of another, $\pi-\pi$ interactions have been the subject of numerous theoretical investigations[36−38] all of which indicate softness of the intermolecular potential as the interaction is lengthened. Allowing an energy range of 1 kcal/mol would permit the distance between the interacting groups to vary from 3.25 to 4.25 Å. This corresponds to a variation of 0.5 Å in the position of the participating ligand atoms with respect to the receptor.

More dramatic are variations along the interaction plane parallel to the face-bound aromatic systems; slippages in nonpolar aromatic systems up to 2.0 Å from ideal geometry just barely approach a 1 kcal/mol limit, and slippages of 1.0 Å require an expenditure of only 0.25 kcal/mol. For aromatic systems with large dipole moments (e.g., nitrogenzene) such slippages are reduced but still provide for variations greater than 1.0 Å within a 1 kcal/mol threshold.[36]

2.3. Hydrophobic Interactions. Hydrophobic interactions in protein systems involve a balance of enthalpic and entropic factors. Enthalpic terms involve the interactions between the protein and itself, the protein and a ligand, the protein or ligand and the solvent, and the solvent with itself. For the former two cases, the interaction involves primarily dispersive forces, whereas the latter two involve a composite of dispersive and electrostatic interactions.[11,12] Notwithstanding, at physiological temperatures, change in the entropy of the solvent (water) tends to be the driving force controlling hydrophobic interactions.[11,12] Within this context, hydrophobic stabilizations are proportional to the area of nonpolar groups exposed to solvent.[11,12]

The factors controlling the extent to which nonpolar moieties on the surface can be hidden from solvent have been an area of active research.[39] Working under the premise that a hydrophobic interaction can be defined by the ability of interacting nonpolar moieties to exclude water[40] (thereby allowing the water to be returned to a nonordered, "bulk" state), a reasonable tolerance in the placement of any specific hydrophobic interaction should be related to the size of a small number of water molecules. Assuming a radius of 1.5 Å for a water molecule, water would be excluded (and hence the region hydrophobic) for intermolecular gaps up to twice this radius. Thus, one can reasonably argue that variations as large as 1.5 Å in the positions of hydrophobic groups would still result in solvent-excluded regions.

2.4. Charge Interactions. The attractive interactions between charged moieties in proteins, and between receptors and ligands, exhibit both dispersive and coulombic components, although it is generally accepted that the coulombic interactions are dominant.[20] Intuitively, one would anticipate that such interactions should have very small variations in the interaction distances due to the $1/r$ nature of electrostatic potential. Furthermore, the fact that charged groups must recoup the large cost of their desolvation to participate in other interactions (e.g., secondary/tertiary structure interactions in proteins or molecular recognition interactions between receptor and ligand) imposes a limit on the amount of energy available for geometric deviations.

A study addressing this issue has been reported by Schneider and co-workers.[41] In the case of the complexation between macrocyclic polyphenolates and tetraalkylammonium compounds, they observed a relationship between the free energy of complexation and the interionic distance similar in form to Coulomb's law

$$\Delta G_0 = \frac{q_1 q_2}{\epsilon r}$$

where q_1 is the charge on the macrocyclic host (−4), q_2 is the charge on the complexed tetraalkylammonium cation (+1), and ϵ is a parameter reflecting the local dielectric environment. The free energy of complexation is attractive for methyl, ethyl, *n*-propyl, and *n*-butyl substituents. These substituents span a range of interionic distances r (based upon measurements of CPK molecular models) between 3.4 (tetramethylammonium) to 4.2 Å (tetra-*n*-butylammonium).

The most attractive interaction for the tetraalkylammonium series is 1.46 kcal/mol.[41] While it is difficult to quantify precisely the magnitude of ionic interactions (salt bridges) in general, it is commonly reported that a single univalent salt bridge (i.e., C+ A−) is stabilized by ~1.25 ± 0.25 kcal/mol in water.[20] This establishes an upper bound on the amount of energy that is available for geometric variations of about 1 kcal/mol.

The tetraalkylammonium data exhibits a typical energy for ionic interactions. Based upon the range of distances from 3.4 to 4.2 Å, we estimate a tolerance of 0.4 Å on the position of a charge center in the ligand involved in an interaction with fixed atoms in the receptor.

2.5. Queries Constructed from Crystallographic Data. Crystallographic structures of proteins often serve as the basis for construction of a query (as illustrated in example 2 below). The accuracy with which it is possible to determine atom positions in the crystal is, of course, a lower bound on the proper tolerances in such queries.

Within the Protein Data Bank, about 75% of the structures solved by crystallography have resolutions between 1.5 and 2.5 Å, and 20% have resolutions above 2.5 Å. A resolution of 2.5 Å allows atom positions to be determined to within 0.4 Å and for a resolution of 1.5 Å to within 0.1 Å.[42]

3. FORM OF GEOMETRIC CONSTRAINTS

Using the chemical function definitions of section 1 and a conformer of a molecule from the database, it is possible to identify the locations (if any) at which a function is present in the conformer. Each location can be represented by one or two points in the coordinate frame of the conformer. Each location can be represented by one or two points in the coordinate frame of the conformer. A point might be the position of an individual atom, the centroid of a group of atoms, or the position of a corresponding atom projected from a hydrogen bond donor or acceptor atom. As mentioned above, binding energies are in part determined by whether there is a conformer of the candidate ligand molecule such that these points can be superimposed with corresponding ideal positions specified in the query, which constitutes an abstraction of the receptor. How should this geometric requirement be described?

3.1. Location versus Distance Constraints. The RMS deviation of corresponding ligand and receptor points, minimized over all possible alignments, provides a good measure of superimposability. A much simpler task than computing the RMS deviation is to check constraints on distances between pairs of ligand points, and for this reason many 3D search systems require queries to be posed in these terms. However, distance constraints do not accurately distinguish conformers that superimpose well with ideal positions from those that do not, for the following two reasons.

Problem of Chirality. As mentioned above, distances are the same for a configuration of atoms and its mirror

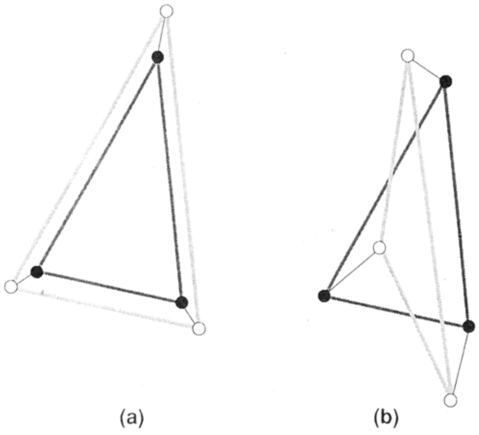


Figure 7. The dark circles represent ideal positions in a query. The open circles represent atom positions. In both (a) and (b), the distances of the query are satisfied to the same degree, as measured by the RMS discrepancy between the lengths of corresponding sides. But the RMS deviation between locations of corresponding points is 6.8 in (a) and much larger, 11.8, in (b). See text for details.

image. Thus, a query with only distance constraints cannot distinguish between two enantiomers, even when one enantiomer superimposes well on the ideal positions and one does not. This problem causes, on average, one additional invalid search hit to be returned for every correct one found.

Problem of Approximation. When there are only three points in the query or when the molecules in the database are achiral or have unknown stereochemistry, the problem of chirality does not apply. However, even in this case distance constraints are a poor approximation to superimposability. For example, both parts a and b of Figure 7 show three ideal positions from a hypothetical query and three corresponding atom positions from a hypothetical conformer, aligned so as to minimize RMS deviation between corresponding points:

$$\sqrt{\frac{1}{N}\sum_{1\le i\le N}(\mathrm{dist}(Q_i,A_i))^2}$$

where N is the number of pairs (3 in this case), Q_i are the ideal positions in the query, and A_i are the atom positions. In both figures, the discrepancy in distances is the same, as measured by the RMS difference between the lengths of corresponding sides of the superimposed triangles:

$$\sqrt{\frac{1}{N(N-1)/2}\sum_{1\le i<j\le N}(\mathrm{dist}(Q_i,Q_j)-\mathrm{dist}(A_i,A_j))^2}$$

However, the RMS deviation between corresponding points is 1.75 times larger in Figure 7b than in Figure 7a.

One can gain some intuition about why this happens from the example in Figure 8. Suppose we want to contrain point C to within some tolerance of a specified location relative to fixed points A and B. If we do this by means of distance constraints, we are forced to include the four regions around the circle as well as the desired circular region.

As an alternative to distance constraints, we propose the concept of location constraints. A location constraint is characterized by a position in 3D space and a tolerance. The position is the ideal location in the receptor for the ligand point being constrained. The tolerance represents an allowable deviation of the actual location from the ideal. When performing a search using location constraints, two requirements are checked:

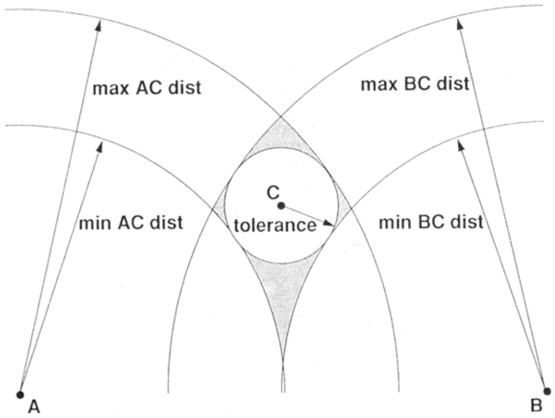


Figure 8. A query intended to find all points within a given tolerance of ideal position C must include in addition the shaded regions if it uses only constraints on distance.

R1. The RMS deviation must be less than one tolerance, i.e.,

$$\sqrt{\frac{1}{N}\sum_{1\le i\le N}\left(\frac{\mathrm{dist}(Q_i,A_i)}{\mathrm{tolerance}_i}\right)^2} < 1$$

R2. The distance between each pair of constrained points must be such that it is possible to align them within the specified tolerances. That is, if the distance between the two ideal positions is d_{12}, and the tolerances are t_1 and t_2, the actual distance d must satisfy $d_{12} - (t_1 + t_2) \le d \le d_{12} + (t_1 + t_2)$. (This requirement is similar to one ordinary distance constraint for each pair of location constraints.)

Note that if it is possible to align the ligand points within the specified tolerances of the ideal positions, both requirements are necessarily satisfied.

The following example shows the difference between distance and location constraints due to the problem of approximation. We use a query for angiotensin II antagonism consisting of three hydrophobes, a hydrogen bond acceptor, and a negative charge center. (The query is fully described in example 1 below). We define a location constraint for each of the seven points: one for each hydrophobe, one for the centroid of the negative group, one for the acceptor atom, and one for the projected position of the corresponding donor atom in the receptor. Searching the BioByteMasterFile database we find 83 hits.

The best equivalent to this location constraint query using distance constraints would be to define a constraint between each pair of points restricting the distance to the ideal value plus or minus the sum of the relevant tolerances (i.e., continue to check R2 but ignore R1). Using such a query we get 155 hits: all 83 found by the original version plus 72 that are an artifact of constraining only distances. If, in an effort to reduce the false positives, the distance constraints are tightened by 10%, we start missing valid hits found by the location constraint query. Tightening the distance constraints is thus not a useful solution.

Table 2 lists the distance constraints and the actual distance values for a conformer of SKF104353Z2 that satisfies the distance version, but not the location version, of the query.

Table 2. Satisfaction of Angiotensin II Antagonism Distance Query by a Conformer of SKF104353Z2[a]

distance between	and	min.	actual	max
hydrophobe 1	hydrophobe 2	5.78	5.89	12.18
hydrophobe 1	hydrophobe 3	8.25	11.39	14.65
hydrophobe 1	acceptor atom	6.08	7.55	12.48
hydrophobe 1	projected point	8.15	9.12	15.75
hydrophobe 1	negative charge center	2.18	4.45	6.38
hydrophobe 2	hydrophobe 3	7.68	8.59	14.08
hydrophobe 2	acceptor atom	2.95	5.56	9.35
hydrophobe 2	projected point	4.54	8.25	12.14
hydrophobe 2	negative charge center	6.11	6.25	10.31
hydrophobe 3	acceptor atom	2.14	5.31	8.54
hydrophobe 3	projected point	1.93	4.52	9.53
hydrophobe 3	negative charge center	9.92	13.47	14.12
acceptor atom	negative charge center	7.93	8.71	12.13
projected point	negative charge center	10.31	11.08	15.71

[a] Values in Å.

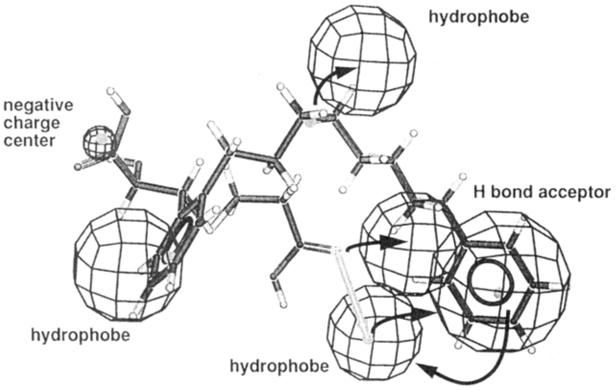


Figure 9. The mesh spheres are the location constraints of the angiotensin II antagonism query. A conformer of SKF104353Z2 which satisfies the distance constraint version of the query is shown optimally superimposed. The arrows highlight features in the compound that are poorly aligned with the specified locations.

The structure of the molecule is shown in Figure 9 along with the alignment of the conformer to the location constraints of the original query. It is clear that the conformer superimposes quite poorly on the ideal locations, despite the fact that it satisfies the distance constraints.

The BioByteMasterFile database does not specify a particular chirality for chiral centers, and so the conformational models cover all possible stereoisomers and, in particular, both enantiomers of each possible conformation. For this reason the above experiment highlights only the effect of the problem of approximation, not the problem of chirality. On a database with specified chiralities, the latter effect would further compound the problem of false positives.

3.2. Other Advantages of Location Constraints. We now consider some other advantages of location constraints. It is often important for queries to exclude atoms of the compound from certain specified volumes. It is possible to use internal constraints (such as distances) to position the excluded volumes with respect to three or more ligand atoms, but this sometimes gives rise to inconsistencies.[1] Let us illustrate this with a familiar example. Figure 10 shows the CNS pharmacophore proposed by Lloyd and Andrews[22] as it would be described in the MOL query language.[6] The query contains a nitrogen, a ring, and excluded volumes, with internal geometric constraints such as distances and angles among them. Suppose that the nitrogen atom lies at its ideal position based on receptor geometry, but that the ring is slightly off. Note that because the excluded volume positions are defined relative to the ring atom positions, they will move with respect to the nitrogen atom, and hence the receptor, as the legal angles between the ring and the nitrogen are explored. It would be more nearly correct for the excluded volumes to be defined with respect to the ideal positions (based on the receptor) rather than the actual positions (based on the ligand) of the nitrogen and ring centroid.

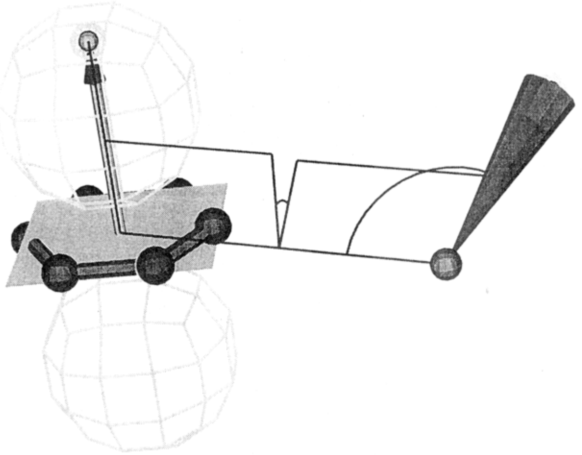


Figure 10. The CNS query of Lloyd and Andrews expressed using internal constraints. The black mesh excluded volume spheres are positioned along the normal to the plane of the phenyl ring.

Table 3. Numbers of Constraints versus Points in Query

points in query	2	3	4	5	6	7	8
location constraints	2	3	4	5	6	7	8
distance constraints (max possible)	1	3	6	10	15	21	28
distance constraints (min. for rigidity)	1	3	6	10	14	18	22

One solution to this problem is to define a set of special "pattern" and "outrigger" points; the ligand atoms satisfying the distance constraints are aligned to the pattern points, and the outrigger points then indicate the location of the excluded volumes.[2] If location constraints are used, they directly fulfill the role of the pattern points, providing a convenient frame of reference for positioning excluded volumes.

Another advantage of location constraints is greater convenience in editing or viewing 3D queries. Location constraints are more easily visualizable than distances, in part because there are fewer of them, especially for complex queries. If there are N points in the query, there can be as many as $N(N - 1)/2$ distances among them. At least four distances must be specified for each point after the fourth to fix its location relative to the previous points. These formulas are compared in Table 3.

Furthermore, a query with a perfectly reasonable looking set of distance ranges might be impossible to satisfy (in three dimensions, at least). Even if it is, only parts of the ranges might be feasible. For example, the three-point query with dist(A,B) = 2−3, dist(B,C) = 2−3, and dist(A,C) = 5−7 can only be satisfied if dist(A,C) = 5−6. What looks like a loose query might actually be very tight or even unsatisfiable.

EXAMPLES

Example 1: Angiotensin II Antagonism. Activity data for 28 compounds was collected from the literature.[43] The data were analyzed using the Catalyst/Hypo system,[15] and a

Table 4. Angiotensin II Antagonism Query[a]

function	tolerance	x	y	z
hydrophobe 1	1.6	0.09	−2.00	0.69
hydrophobe 2	1.6	6.75	3.63	−1.48
hydrophobe 3	1.6	9.10	−6.46	−4.79
negative charge center	0.5	−0.64	0.31	−2.84
acceptor atom	1.6	9.04	−2.07	−1.75
projected point	2.2	11.85	−2.96	−1.18

[a] Values in Å.

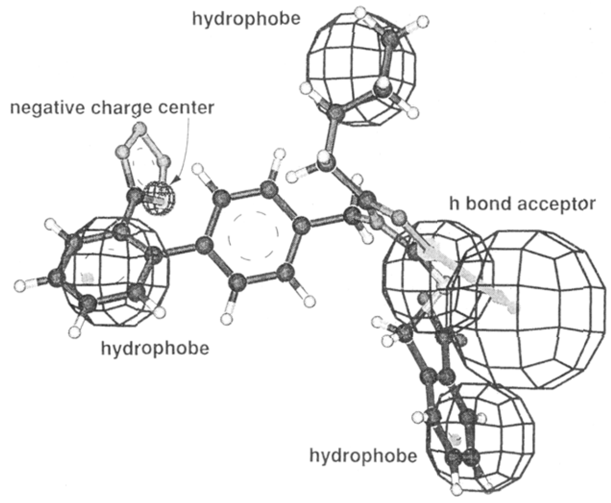


Figure 11. Compound **71** of ref 8 superimposed on angiotensin II antagonism query.

structure−activity hypothesis was produced that accounted for the data. The corresponding query has five chemical functions as shown in Table 4: three hydrophobic regions, a negative charge center, and a hydrogen bond acceptor. The query and its generation are reported elsewhere.[44] A set of new active compounds was later published by Ashton et al. of Merck Research Laboratories.[8] These compounds were not known to us at the time we proposed our model but were found to fit it very well. Figure 11 shows our query mapped to the triazole compound 71 reported by Merck. This mapping and the nature of the interactions (lipophilic, H-bond, and basic sites[8]) are consistent with the model postulated by the Merck group.

Searching with this query identifies 83 compounds in the BioByteMasterFile, 9 in Maybridge, 86 in NCI, and 737 in Derwent. Note that because stereochemistry is largely unspecified in these databases, a molecule whose enantiomer satisfied the query could not be ruled out by the search (as is the case when stereochemistry is known).

Of the 19 compounds in Derwent that best fit the query, seven are identified in the database as known angiotensin II antagonists. Among these were X-6803, which uses a carboxylate to satisfy the negative charge center feature, and L-158809, which uses a tetrazole nitrogen. These molecules, shown in Figure 12, demonstrate the value of the generalized charge center definition: if the query described just the tetrazole nitrogen typical of most published angiotensin II antagonists, it would have missed other compounds known to be active.

Example 2: Human Leukocyte Elastase Inhibitor. The X-ray crystal structure for a peptidic inhibitor bound to human neutrophil elastase, a hydrolase for elastin, was recently reported at a resolution of 1.84 Å.[45] The bound inhibitor is methoxysuccinyl-Ala-Ala-Pro-alanine*, where the alanine* residue has had the carboxy carbonyl replaced by a methylene group. It is precisely this carbonyl group that would be attacked by Ser 214.

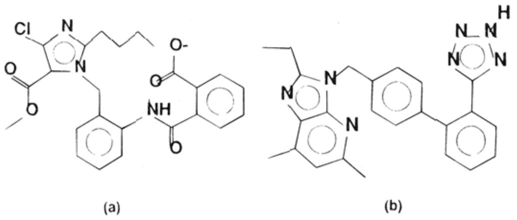


Figure 12. Two known active compounds found using angiotensin II antagonism query. X-6803 (a) satisfies the negative charge center with a carboxylate. L-158809 (b) uses a tetrazole.

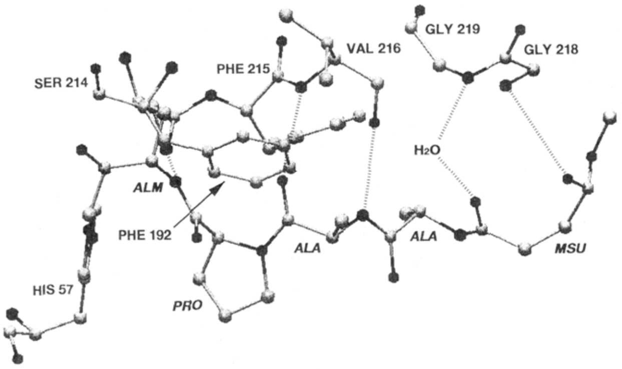


Figure 13. HLE binding schematic.

Inspection of the bound inhibitor structure reveals that the primary interaction between the protein and substrate involves a total of five hydrogen bonds, shown in Figure 13. These hydrogen bonds involve the carbonyl oxygens of Ser 214 and Val 216, and the peptide NHs of Val 216, Gly 218, and Gly 219. The hydrogen bond interaction involving the NH of Gly 219 is mediated by an interstitial water. Also, the pseudo N-terminal end of the inhibitor is more loosely bound as the natural substrate would require space for a larger Val-Gly-Val subunit. For this reason we choose to include only the other three hydrogen bonds in our query.

A secondary interaction that appears to have some importance involves the proline ring and the three residues that flank it. While there is no defined hydrophobic "pocket", Phe 192, Phe 215, and the imidazole ring of His 57 provide a hydrophobic environment that nicely accommodates the proline ring.

Using the crystal structure as a starting point, a query can be defined using generalized chemical functional definitions. Two hydrogen bond donors, one hydrogen bond acceptor, and one hydrophobe are employed. The two hydrogen bond donors on the inhibitor are placed to interact with the acceptors on Ser 214 and Val 216. Between these two donors is a hydrogen bond acceptor situated to interact with the donor on Val 216. The hydrophobe position is chosen to model the interaction involving the proline ring. The tolerances for each of these interactions are chosen with consideration of the experimental data discussed above and are summarized in Table 5. The use of a larger tolerance for the projected donor/acceptor points reflects an attempt to combine the distance and angular variations in hydrogen bonding interactions.

Excluded volume spheres are added at four locations to reflect the steric requirements of the receptor. Ser 214, the

Table 5. Human Leukocyte Elastase Inhibitor Query[a]

function	tolerance	x	y	z
donor atom 1	1.6	1.97	2.39	−0.67
projected point	2.2	0.98	0.40	1.22
donor atom 2	1.6	6.74	3.12	2.95
projected point	2.2	7.67	0.47	2.41
acceptor atom	1.6	3.78	2.48	1.47
projected point	2.2	2.87	−0.12	2.91
hydrophobe	1.6	6.96	4.17	−0.92
excluded volume sphere	3.0	−4.64	0.32	1.32
excluded volume sphere	2.5	10.20	−2.84	−0.27
excluded volume sphere	3.0	3.70	5.00	6.00
excluded volume sphere	3.0	3.70	−2.00	−3.00

[a] Values in Å.

residue responsible for the enzyme's hydrolytic activity, and the phenyl rings of Phe 192 and Phe 215 are each represented with a single exclusion sphere (three in total). The interstitial water that participates in a hydrogen bond to the methoxy succinyl moiety is also modeled by a single excluded volume. The radius of each exclusion sphere is chosen so as to mimic closely the overall shape of the receptor components mentioned.

Searching with this query identifies 1710 compounds in the BioByteMasterFile, 1110 in Maybridge, 4035 in NCI, and 5161 in Derwent.

CONCLUSIONS

The key characteristics of chemically reasonable 3D database queries have been considered. Chemists are searching for novel structures that can bind to an enzyme or receptor. A query describing the chemical functions involved in binding is more likely to capture this intent than a query based on specific atomic topologies. Generalized functional definitions for hydrogen bond acceptors and donors, charge centers, and hydrophobic regions have been proposed. For the 5-HT_3 pharmacophore of Hibert et al., 76% of the compounds found with a functional query are missed with the original topological query. For the ACE pharmacophore of Sprague, 70% of the hits found are missed if an aromatic ring feature is substituted for a generalized hydrophobe. In both cases, the functional definitions also eliminate false positives where key features were not surface accessible.

The tolerance in a geometric constraint of a query must make physical-chemical sense. It has been shown that the following variations in position are reasonable for energies likely to be encountered in bound structures.

hydrogen bonding atoms: 0.5−2.1 Å
π−π interactions: 0.5−2.0 Å
hydrophobic interactions: ~1.5 Å
charge interactions: ~0.4 Å

For instance, the actual position of a charge center in the ligand could be anywhere within 0.4 Å of the ideal position and still permit a significant interaction. Note that the tolerance on a distance between two points would be the sum of the positional tolerances shown above for the two points. While the above values are of course approximate, it does seem clear that tolerances as tight as 0.1 Å[7] or even 0.006 Å[4] on distances would be hard to justify in practical queries.

It has also been shown that internal constraints on distances between ligand atoms cannot eliminate conformers that superimpose poorly with ideal positions, which is what matters for binding. It is worth noting that most sources of information a chemist has about the class of molecules being sought provide location rather than distance information. This is true for X-ray structures, pharmacophores obtained by mutual superposition, CoMFA coefficient maps, and structure−activity hypotheses. Thus, conversion of data from these sources into distance constraints necessarily entails some loss of information. For these reasons, constraints on location are often preferable.

Putting all of these findings together leads us to the following observation. Queries used in the literature to benchmark database search algorithms[4,6,7] typically refer to specific atomic topologies with very tight distance constraints among only three to five points. More realistic queries would use generalized functions instead of atomic topologies and larger, physically reasonable tolerances. Adequate selectivity should be maintained instead by adding more features to the query, verifying their surface accessibility, constraining locations rather than just distances, and adding included or excluded volumes where indicated. A limit on the energy of the conformer satisfying the query also helps in this regard.

ACKNOWLEDGMENT

The authors gratefully acknowledge the crucial contributions of Shneor Berezin and Siuling Ku (search engine), Gary Chappell (user interface), Andrew Smellie and Peter Towbin (generation of conformational models), and John Van Drie (function dictionary concept), all of whom helped make this study possible. A share of credit (but none of the blame) for the chemical function definitions goes to Professor Paul Bartlett (University of California, Berkeley), Frank Brown (Glaxo), Chris Culberson (Merck), and Mark Hermsmeier (Bristol-Myers Squibb), who provided comments on early versions. Rémy Hoffman and Daniel Parish are thanked for providing valuable insights during the design and testing of the search system and function dictionary. Finally, we acknowledge Silicon Graphics Inc. for helping to provide the computing resources needed throughout this work.

APPENDIX: HYDROPHOBE IDENTIFICATION ALGORITHM

Given a conformer, we want to identify its hydrophobic surface regions and to mark the effective center of each region with a point. Before describing the algorithm, we need a few definitions.

The accessible surface s of an atom is the portion of the surface that can be touched by a probe sphere that does not collide with any other atom of the molecule. We find that the following fast, approximate method for calculating s suffices in practice. A set of points is distributed uniformly on the surface of the atom in question. A probe sphere is placed tangent to the atom at each point, and we check whether it collides with the van der Waals radius of any other atom of the molecule. We determine the fraction f of the grid points that are accessible and set $s = 4\pi r^2 f$ where r is the van der Waals radius of the atom.

The topology-dependent term t for an atom depends on the type of the atom and its neighbors. Because our goal is to assign hydrophobicity to small groups of atoms, our typing system works differently from, say, that of Ghose and Crippen[23] where the goal is to reproduce a partition coefficient for the entire molecule. In particular, we reduce the

Table 6. Topology-Dependent Hydrophobicity Factors

category	factor	description
1	0	N, O, or H
2	0	S in SH
3	0	≤2 bonds away from charged atom
4	0	≤2 bonds away from OH or NH with no delocalized electrons
5	0	≤1 bond away from SH with no delocalized electrons
6	0	≤2 bonds away from O with double bond
7	0	≤1 bond away from S with valence > 2
8	0	S with double bond
9	0.6	3 bonds away from O with double bond
10	0.6	2 bonds away from S with valence > 2
11	0.6	1 bond away from S with double bond
12	0	two or more instances of any of the previous three conditions
13	0.25	1 neighboring O or N with no delocalized electrons
14	0	>1 neighboring O or N with no delocalized electrons

hydrophobicity of an atom if there are nearby hydrophilic atoms rather than simply adding an opposing contribution for the hydrophilic atoms. Our t values are determined according to a set of simple rules which were chosen to reflect the judgment of medicinal chemists on a set of test molecules as succinctly as possible. Hydrogens are ignored; their contribution is reflected in the value of the adjacent atom. The t value for a non-hydrogen atom is 1 times the appropriate factors for all of the categories into which the atom falls, as defined in Table 6. For example, all three atoms in CC=O fall in category 6 ("≤2 bonds away from O with double bond") and so get $t = 0$. A negatively charged O falls in category 3 and so gets $t = 0$. C in C≡N falls in category 13 and so gets $t = 0.25$. Finally, Cl in CH_3Cl falls in no special category, and so gets a full $t = 1$.

The hydrophobicity of an atom is $h = ts$.

Let $h_{\min}$ be one half the h value of an exposed methyl carbon terminating a carbon chain. We require that each identified hydrophobic group include at least this much hydrophobicity.

Once we have determined h for each atom, we proceed to identify groups of atoms which form a hydrophobic region as follows:

1. Define groups for rings of size 7 or less. A group is defined for each such ring satisfying the following conditions:

   a. the sum of h values for the atoms in the ring is at least $h_{\min}$, and

   b1. all substituents are on one side of the plane of the ring, or

   b2. at least two neighboring ring atoms have $h > 0$ and no substituent of more than two atoms.

(The intent of conditions b1 and b2 is to be sure that the ring presents one large surface rather than two or more small ones.) Each group is marked by the centroid of the ring's atoms, weighting each atom by its h value. Then all atoms in rings of size 7 and smaller are removed from further consideration.

2. Define groups for atoms with three or more bonds. A group is defined for each atom with three or more bonds, and those of its neighbors that are not bonded to any other atom, provided that the sum of the h values is at least $h_{\min}$. Each group is marked by the centroid of the atoms, weighting each atom by its h value. Then all atoms with three or more bonds and all of their neighbors having only one bond are removed from further consideration.

3. Divide the remaining atoms into chains. First, all atoms with $h = 0$ are removed. Any rings that are left can be treated like chains by arbitrarily choosing one ring atom to be the start of a chain.

4. Define zero or more groups for each chain. Chains are divided into contiguous groups such that the sum of h values for the atoms in each group is at least $h_{\min}$ and less than twice $h_{\min}$. If exactly one of the atoms in the group has only one bond, the group is marked by that atom's position. Otherwise the group is marked by the weighted centroid of the atoms.

REFERENCES AND NOTES

(1) Van Drie, J.; Weininger, D.; Martin, Y. ALADDIN: An integrated tool for computer-assisted molecular design and pharmacophore recognition from geometric, steric, and substructure searching of three-dimensional molecular structures. *J. Comp.-Aided Mol. Design* **1989**, *3*, 225–251.

(2) Sheridan, R.; Nilakantan, R.; Rusinko, A., III; Bauman, N.; Haraki, K.; Venkataraghavan, R. 3DSEARCH: A system for three-dimensional substructure searching. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 255–260.

(3) Bartlett, P. A.; Shea, G. T.; Telfer, S. J.; Waterman, S. CAVEAT: A Program to Facilitate the Structure-derived Design of Biologically Active Molecules. In *Molecular Recognition*: *Chemical and Biological Problems*; Roberts, S. M., Ed.; Royal Society of Chemistry: London, 1989; Vol. 78, pp 182–196. Christie, B. D.; Henry, D. R.; Guner, O. F.; Moock, T. E. MACCS-3D: A tool for Three-dimensional Drug Design. In *Proceedings of the 14th International Online Information Meeting*; Raitt, D., Ed.; Learned Information: Oxford, U.K., 1990; pp 137–161. Murrall, N. W.; Davies, E. K. Conformational Freedom in 3-D Databases. 1. Techniques. J. *Chem. Inf. Comput. Sci.* **1990**, *30*, 312–316. Clark, D.; Willett, P.; Kenny, P. Pharmacophoric pattern matching in files of three-dimensional chemical structures: use of bounded distance matrices for the representation and searching of conformationally flexible molecules. *J. Mol. Graphics* **1992**, *10*, 194–204.

(4) Hurst, T. Flexible 3D Searching: The Directed Tweak Technique. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 190–196.

(5) Martin, Y. 3D Database Searching in Drug Design. *J. Med. Chem.* **1992**, *35*, 2145–2154.

(6) MDL Information Systems. 3D Searching Strategies. 1. Formulation of Pharmacophoric Queries. MACCS-3D Applications Technical Notes 1991.

(7) Clark, D.; Jones, G.; Willett, P.; Kenny, P.; Glen, R. Pharmacophoric Pattern Matching in Files of Three-Dimensional Chemical Structures: Comparison of Conformational-Searching Algorithms for Flexible Searching. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 197–206.

(8) Ashton, W. et al. Nonpeptide Angiotensin II Antagonists Derived from 4*H*-1,2,4-Triazoles and 3*H*-Imidazo[1,2-*b*][1,2,4]triazoles. *J. Med. Chem.* **1993**, *36*, 591–609.

(9) Hibert, M.; Hoffmann, R.; Miller, R.; Carr, A. Conformation–Activity Relationship Study of 5-HT_3 Receptor Antagonists and a Definition of a Model for this Receptor Site. *J. Med. Chem.* **1990**, *33*, 1594–1599.

(10) Bush, B.; Sheridan, R. PATTY: a programmable atom typer and language for automatic classification of atoms in molecular databases. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 756–762.

(11) Privalov, P. L.; Gill, S. J. Stability of Protein Structure and Hydrophobic Interaction. *Adv. Protein Chem.* **1988**, *39*, 191–234.

(12) Dill, K. A. Dominant Forces in Protein Folding. *Biochemistry* **1990**, *29*, 7133–7155.

(13) Martin, Y.; Bures, M.; Danaher, E.; DeLazzer, J. New strategies that improve the efficiency of the 3D design of bioactive molecules. *Trends in QSAR and Molecular Modelling 92*, Wermuth, C., Ed.; ESCOM: Leiden, The Netherlands, 1993; pp 20–26.

(14) Rigby, M.; Smith, E. B.; Wakeham, W. A.; Maitland, G. C. *The Forces Between Molecules*; Clarendon Press: Oxford, U.K.; 1986.

(15) *Catalyst 2.2 Tutorial Manual.* Molecular Simulations, Inc.: Burlington, MA, 1994.

(16) Berezin, S.; Greene, J.; Kahn, S.; Ku, S.; Teig, S. CHM: A New Chemically Expressive Database Query Language. Noordwijkerhout Camarino Medicinal Chemistry Symposium; The Netherlands, 1993.

(17) The BioByteMasterFile, NCI, Derwent and Maybridge databases are available from Molecular Simulations, Inc., Burlington, MA.

(18) Smellie, A.; Teig, S. L.; Towbin, P. Poling: Promoting Conformational Variation. *J. Comput. Chem.* In press.

(19) Brooks, B.; Bruccoleri, R.; Olafson, B.; States, D.; Swaminathan, S.; Karplus, M. CHARMm: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **1983**, *4*, 187−217.

(20) Schneider, H.-J. Mechanisms of Molecular Recognition: Investigations of Organic Host−Guest Complexes. *Angew. Chem., Int. Ed. Engl.* **1991**, *30*, 1417−1436. Also references contained therein.

(21) Roberts, S. M. *Symposia in Print*: *Molecular Recognition in Chemistry and Biochemistry Problems*; Royal Society: London, 1989.

(22) Lloyd, E.; Andrews, P. A common structural model for central nervous system drugs and their receptors. *J. Med. Chem.* **1986**, *29*, 453−462.

(23) Ghose, A.; Crippen, G. Atomic Physicochemical Parameters for Three-Dimensional Structure-Directed Quantitative Structure−Activity Relationships I. Partition Coefficient as a Measure of Hydrophobicity. *J. Comput. Chem.* **1986**, *7*, 565−577.

(24) Eisenberg, D.; McLachlan, A. Solvation Energy in Protein Folding and Binding. *Nature* **1986**, *319*, 199−203.

(25) Fauchere, J.-L.; Quarendon, P.; Kaetterer, L. Estimating and Representing Hydrophobicity Potential. *J. Mol. Graphics* **1988**, *6*, 203−206.

(26) Furet, P.; Sele, A.; Cohen, N. 3D molecular lipophilicity potential profiles: a new tool in molecular modeling. *J. Mol. Graphics* **1988**, *6*, 182−189. Suzuki, T.; Kudo, Y. Automatic log P estimation based on combined additive modeling methods. *J. Comp.-Aided Mol. Design* **1990**, *4*, 155−198. Croizet, F.; Langlois, M.; Dubost, J.; Braquet, P.; Audry, E.; Dallet, P.; Colleter, J. Lipophilicity Force Field Profile: An Expressive Visualization of the Lipophilicity Molecular Potential Gradient. *J. Mol. Graphics* **1990**, *8*, 153−155.

(27) Lee, B.; Richards, F. The Interpretation of Protein Structures: Estimation of Static Accessibility. *J. Mol. Biol.* **1971**, *55*, 379−400.

(28) Petrillo, E.; Ondetti, M. Angiotensin-Converting Enzyme Inhibitors: Medicinal Chemistry and Biological Actions. *Med. Res. Rev.* **1982**, *2*, 1−41.

(29) Sprague, P. Building a Hypothesis for Angiotensin Coverting Enzyme Inhibition. Catalyst Application Note. BioCAD Corp.: Sunnyvale, CA, 1994.

(30) Murray-Rust, P.; Glusker, J. Directional Hydrogen Bonding to sp^2- and sp^3-Hybridized Oxygen Atoms and Its Relevance to Ligand-Macromolecule Interactions. *J. Am. Chem. Soc.* **1984**, *106*, 1018−1025.

(31) Taylor, R.; Kennard, O.; Versichel, W. Geometry of the N−H··O=C hydrogen bond. 1. Lone pair directionality. *J. Am. Chem. Soc.* **1983**, *105*, 5761−5766. Taylor, R.; Kennard, O.; Versichel, W. Geometry of the N−H··O=C hydrogen bond. 3. Hydrogen-Bond Distances and Angles. *Acta Crystallogr.* **1984**, *B40*, 280−288.

(32) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A Geometric Approach to Macromolecular-Ligand Interactions. *J. Mol. Biol.* **1982**, *161*, 269−288.

(33) Boobbyer, D.; Goodford, P.; McWhinnie, P.; Wade, R. New hydrogen-bond potentials for use in determining energetically favorable binding sites on molecules of known structures. *J. Med. Chem.* **1989**, *32*, 1083−1094.

(34) Baker, E. N.; Hubbard, R. E. Hydrogen Bonding in Globular Proteins. *Prog. Biophys. Mol. Biol.* **1984**, *44*, 97−179.

(35) Burley, S. K.; Petsko, G. A. In *Advances in Protein Chemistry*; Anfinsen, C. B., Edsall, J. T., Richards, F. M.; Eisenberg, D. S., Eds.; Academic Press: 1988.

(36) Hartsough, D. S. Ph.D. Thesis, University of Illinois at Urbana-Champaign, 1991.

(37) Hunter, C. A.; Saunders, J. K. M. *J. Am. Chem. Soc.* **1990**, *112*, 5525.

(38) Karlström, G.; Linse, P.; Wallqvist, A.; Jönsson, B. Intermolecular Potentials for the H_2O−C_6H_6 and the C_6H_6−C_6H_6 Systems Calculated in an ab Initio SCF CI Approximation. *J. Am. Chem. Soc.* **1983**, *105*, 3777−3782.

(39) Connolly, M. L. *J. Am. Chem. Soc.* **1985**, *107*, 1118−1124.

(40) Dean, P. M. *Molecular Foundations of Drug-Receptor Interaction*; Cambridge University Press: Cambridge, 1987; p 104.

(41) Schneider, H.-J.; Güttes, D.; Schneider, U. Mechanisms of Molecular Recognition: Investigations of Organic Host-Guest Complexes. *J. Am. Chem. Soc.* **1988**, *110*, 6449−6454.

(42) Fersht, A. *Enzyme Structure and Mechanism*, 2nd ed.; W. H. Freeman and Co.: New York, 1985; p 6.

(43) Carini, D. et al. *J. Med. Chem.* **1990**, *33*, 1330−1336. Duncia, J. et al. *J. Med. Chem.* **1990**, *33*, 1312−1329. Weinstock, J. et al. *J. Med. Chem.* **1990**, *34*, 1514−1517. Mantlo, N. et al. *J. Med. Chem.* **1990**, *34*, 2919−2922. Buhlmayer, P. et al. *J. Med. Chem.* **1990**, *34*, 3105−3114. Bradbury, R. et al. *J. Med. Chem.* **1990**, *35*, 4027−4038. De, B. et al. *J. Med. Chem.* **1990**, *35*, 3714−3717.

(44) Sprague, P. Building a Hypothesis for Angiotensin II Antagonism. Catalyst Application Note. BioCAD Corp.: Sunnyvale, CA, 1994.

(45) Navia, M. A.; McKeever, B. M.; Springer, J. P.; Lin, T.-Y.; Williams, H. R.; Fluder, E. M.; Dorn, C. P.; Hoogsteen, K. *Proc. Natl. Acad. Sci. U.S.A.* **1989**, *86*, 7−11.