Use of Flexible Queries for Searching Conformationally Flexible Molecules in Databases of Three-Dimensional Structures

OSMAN F. GÜNER,*,† DOUGLAS R. HENRY,† and ROBERT S. PEARLMAN*,‡

Molecular Design Limited, 2132 Farallon Drive, San Leandro, California 94577, and College of Pharmacy, University of Texas at Austin, Austin, Texas 78712

Received July 20, 1991

Searching for conformationally flexible molecules is currently a topic of widespread study. In general, this type of three-dimensional (3D) search can proceed by locating the flexibility information either in a database (by storing multiple conformers or conformational analysis results), or in a search engine (by implementing flexible fitting methods), or in a search query (by combining fixed and mobile query features). This paper describes the third approach, using a flexible pharmacophoric query for ACE inhibitors. 3D searches were performed using the MACCS-3D structural database system to search MDDR-3D, a database of about 17 000 3D drug structures of current interest. The structures in the database were built using the CONCORD program. The paper describes a stepwise procedure for building the query, performing searches, and optimizing the query to obtain a high ratio of structures with the desired activity. The results offer a useful solution to the problem of searching conformationally flexible molecules in databases of 3D structures and illustrate how to introduce conformational flexibility in the domain of the search query.

INTRODUCTION

Three-dimensional (3D) searching has emerged as a new and particularly promising approach to computer-assisted molecular design. While other applications are certainly of interest (e.g., exploration of structure/property relationships), an increasingly useful application involves searching large databases of 3D chemical structures for those compounds which satisfy both the chemical and geometric criteria necessary for favorable interaction with a biomolecular receptor. A 3D search query is formulated based upon what would be called, in pharmaceutical terms, a pharmacophore map—an indication of the chemical nature of binding subsites within the receptor site and an indication of the relative spatial positions and orientations of those subsites. A 3D search query reflects the hypothesized chemistry and geometry of potentially bioactive compounds in their receptor-bound conformations.

Several reviews[1,2] and articles[3-6] have indicated that 3D searching, as currently implemented in various software packages, works quite well. However, some fundamental improvements would make the approach even more useful. One set of improvements addresses what has come to be known as "the conformational flexibility issue" (vide infra).

A 3D search query is intended to reflect the receptor-bound conformation of potentially bioactive compounds. However, the 3D structures stored in databases are generated without knowledge of how the receptor structure and binding process might affect the conformation of the ligand. Instead, databases often contain only a single, low-energy conformation for each compound. If a stored conformation of a particular flexible compound differs significantly from the bound conformation and if the search query does, indeed, reflect the bound conformation, then a search with that query would fail to retrieve that compound. This is the essence of the "conformational flexibility issue".

Three approaches address the issue of conformational flexibility: (i) exploration of the conformational space of selected compounds at search time, i.e., conformationally flexible searching (CFS); (ii) storing a large number of conformations of each compound in the database; and (iii) incorporating flexibility information into the 3D query used to search databases of one or a small number of low-energy conformations.

One may also consider, as a fourth approach, including conformational flexibility information for each entry in the database. Such an approach will require a conformational analysis for each structure before or during database registration; furthermore, it will require modification of the search procedure to incorporate this information. We believe that this method can be a part of CFS (approach no. 1). In fact, early studies indicate beneficial use of a distance-range key-screening process as a prerequisite for CFS. Intuitively, this first approach is very attractive and is currently an important research topic. Research and development efforts in this area are being conducted at a number of sites and therefore will not be covered in this paper.

Arguments against the second approach are compelling. In large 3D databases, storing multiple conformations, even if there are only a few per structure, will quickly overwhelm the disk space and CPU requirements of typical computational facilities. Moreover, this approach fails to address the crux of the conformational flexibility issue: There is no guarantee that the bound conformation corresponds to *any* of the local energy minima identified and stored for the unbound compound.

Meanwhile, the third approach to the conformational flexibility issue (i.e., building conformational flexibility into the search query) represents a reasonable interim solution to the problem. Currently, conformational flexibility is built into the search query by specifying broad ranges of acceptable distances, angles, and dihedral angles between objects (atoms, groups, lone-pairs, ring-centroids) combined in a search query. For example, imagine a search for compounds containing both a hydrogen-bond accepting group and a nitro group. Imagine, also, that the distance between the corresponding receptor subsites requires that these groups be 8.1 Å apart. If the search query demanded that the intergroup distance be precisely 8.1 Å, the number of hits returned after searching a database of rigid, low-energy conformations might be very small. If, instead, the search query accepts all compounds with nitro to H-bond acceptor distances between 6 and 10 Å, we increase the probability of finding compounds which might achieve the desired 8.1-Å intergroup distance through relatively modest conformational distortion. However, this approach is prone to both false negatives and false positives. False negatives result when compounds, for which the intergroup distance falls outside this range, might achieve the 8.1-Å intergroup distance by conformational changes which are more significant in the

† Molecular Design Limited.
‡ University of Texas at Austin.

0095-2338/92/1632-0101$03.00/0

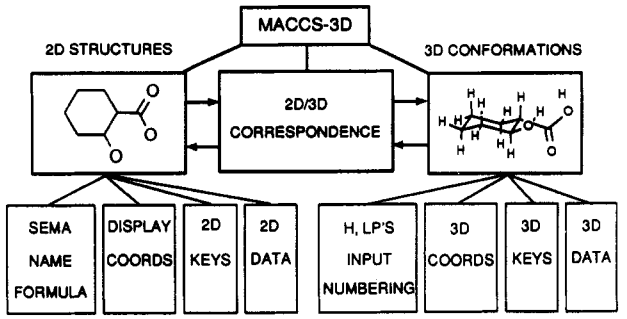


Figure 1. Schematic diagram of MACCS-3D. Structures and conformations have separate data files and internal registry numbers. They are linked through a 2D/3D correspondence file. Each structure may have multiple conformations stored. H's and LP's refer to hydrogens and lone pairs, respectively.

geometric (but not necessarily energetic) sense. False positives result when the search returns compounds for which the desired 8.1-Å intergroup distance cannot actually be achieved. Increasing the specified distance range might decrease the number of false negatives but would increase the number of false positives (and vice versa).

This paper describes an approach that reduces the false negatives and positives and avoids this dilemma: an approach to searching a 3D database of single structures for conformationally flexible compounds using "flexible queries". The idea behind this approach is to develop a query with a firm "grip" (constraint) on the rigid or semi-rigid components (hereafter referred to as rigid throughout the text) of the pharmacophoric groups of a compound while, at the same time, maintaining a relaxed but controlled grip on the flexible components.

METHODS

The searching software used was MACCS-3D[7] (Release 1.0), and the 3D database searched was the MACCS-II Drug Data Report-3D (MDDR-3D) database[8] (Release 90.2), built using CONCORD[9,10] (v 2.9.1). The following paragraphs describe the searching software, database, and method by which the database was generated.

Search Software. The MACCS-3D program, introduced in 1989, is a new module of the MACCS-II System (Molecular ACCess System[11]). MACCS-3D provides three main functions: (1) input and storage of 3D conformations and data, (2) searching using a variety of 3D geometric queries, and (3) output of conformations and data to forms, lists, or molecule files. Figure 1 shows a schematic of the MACCS-3D system.

Conformations are input to MACCS-3D from molecule files, which may also contain per-structure, per-conformation, per-atom, or atom-pair data. The minimum data requirements are a connection table and 3D coordinates. Other information such as atom charge, isotope, and hydrogen count may be included as part of the connection table. Atom and double-bond stereochemistry are perceived automatically by the program from the connection table and coordinates.

Search Process. The search process involves four stages: (1) defining the query, (2) interpreting the search query, (3) 2D and 3D key screening, and (4) atom-by-atom mapping and applying of 3D constraints. A search query in MACCS-3D may include any combination of the following: 2D substructures, 3D objects and constraints, fixed atoms or fragments, and per-atom or atom-pair data constraints. Figure 2 shows the 3D objects and constraints allowed by MACCS-3D, and Figure 3 shows a query with each type of geometric feature represented. Incorporating atom and atom-pair data into geometric searches is discussed elsewhere.[12] In general, each 3D feature can be assigned a unique name, a search tolerance, and for display purposes, a color. The 3D features

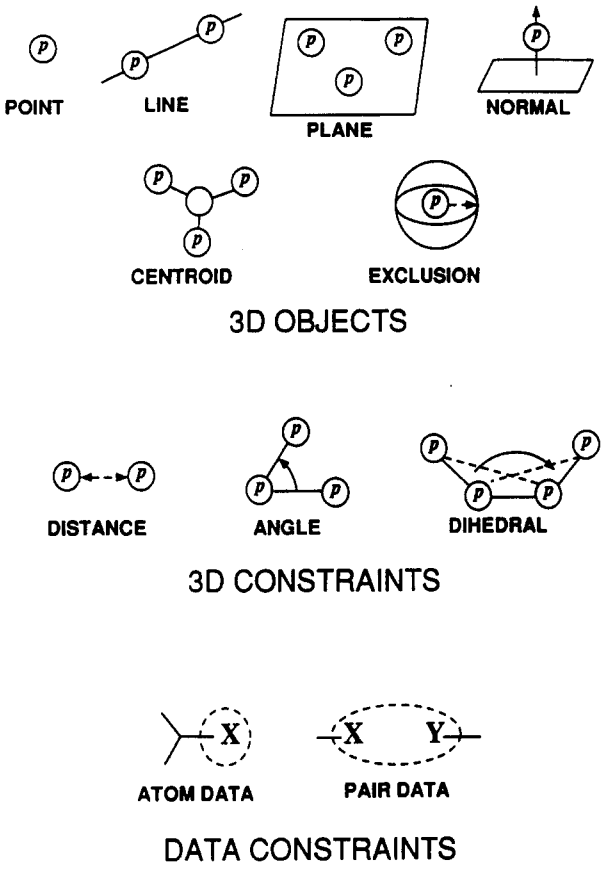


Figure 2. Search query objects, constraints, and data in MACCS-3D. RMS tolerance (Å) can be given for lines and planes; exclusion spheres can be defined by a radius (Å) from a central point; range of distances (Å), angles, and dihedral angles (deg) can be given for the 3D constraints. Points can represent atoms as well as dummy points.

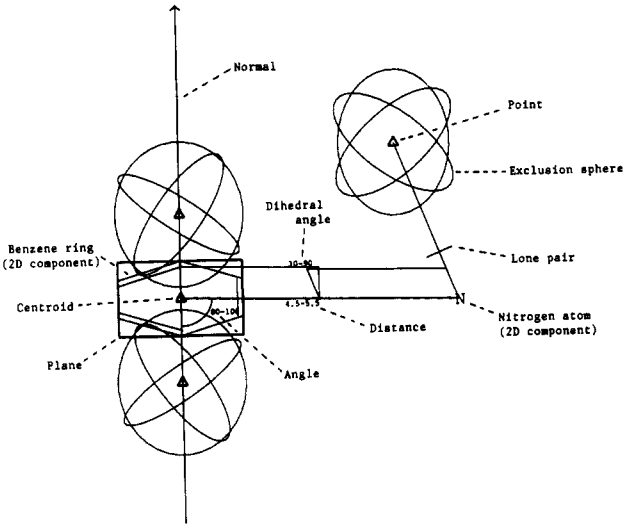


Figure 3. 3D query illustrating many of the 3D query features available in MACCS-3D. This example displays the CNS-active drugs pharmacophore proposed by Lloyd and Andrews.[25]

can be displayed as part of the query, or they may be collected, mapped onto a conformation, and displayed with measured values on a plot of the conformation. A variety of display options can be used to enhance the display, including stereo and perspective displays and scaling and orientation tools.

Interpretation of the search query involves (1) analysis of the atom, bond, and 3D feature types in the query, (2) generating 2D and 3D key screens, and (3) generating a search script. The 2D keys identify the presence of structural features and combinations of atom, bond, and ring types. The 3D keys

mark certain ranges of distances and angles involving atoms in the structure. The search script is a linear sequence of operation codes that are executed during the search process. These instruct the search engine to find certain atoms, bonds, and 3D features in the molecule which meet the search criteria. The search script can be backtracked by the search engine if a "dead end" is reached. Once all the atoms, bonds, and features in the query are matched, the molecule is considered a "hit". The 2D (atom-by-atom mapping) and 3D (geometric features and fixed-atom calculation and comparison) search operations are integrated into the search program. Thus, 2D and 3D searching take place simultaneously in MACCS-3D. This approach differs from some other systems, such as ALADDIN[5,13] and ChemDBS-3D,[14] which conduct a 2D search first to obtain a complete atom-to-atom mapping and then perform the 3D calculations in a following step. Each approach has advantages and disadvantages depending on the nature and the complexity of the search query.

The search process involves first screening against 3D and 2D keys. The resulting hit list is then subjected to 3D substructure searching. This identifies all nonoverlapping subgraph isomorphs which contain the required 3D constraints. The program maps and compares fixed atoms by an overall root-mean-square (RMS) fit, and by a maximum per-atom deviation from the query atom. It also conducts integral per-atom and atom-pair data searches. The result is a list of all structures from the original list which match the search query to the desired degree of tolerance. The list of structures can be viewed, merged with other lists, or saved in a file.

Database. For this study, we used the 3D version of the MACCS-II Drug Data Report Database (MDDR-3D).[8] This database is derived from the Prous journal *Drug Data Report*, which describes bioactive compounds for which patents have been filed recently.[15] The current version (Rev. 90.2) of this database contains single CONCORD-generated 3D structures for 16703 of the 20367 2D structures appearing in *Drug Data Report* between July 1988 and June 1990. The structures span a wide range of compound classes and biological activities. Other information stored includes chemical and generic names, company, trademark, CAS Registry Number, a variety of patent and literature references, physical properties, molecular formula, 3D structure source, and comments related to modeling of the structure. The 3D structures are stored without explicit hydrogens or lone pairs; MACCS-3D can attach these at search time if the search query contains them.

Database Generation. The connection tables (MOLfiles) contained in the MDDR database were broken into single-fragment MOLfiles, processed using CONCORD,[9,10] and reassembled into multi-fragment, 3D MOLfiles. CONCORD is a program for the rapid conversion of 2D or 2.5D CONnection tables to high quality, approximate 3D CoORDinates, and is distributed by Tripos Associates.

The CONCORD program is a hybrid of an "expert system" approach and a "pseudo molecular mechanics" approach to structure building. Bond lengths are assigned from a very extensive table.[16] For most acyclic substructures, the bond angles and torsion angles are assigned by performing a rule-based logical analysis which seeks to optimize 1–4 interactions. This logical analysis is at least 1000 times faster than performing the analogous task using a molecular mechanics approach. This rule-based approach cannot be applied to rings or ring systems; minimization of conformational strain is required. However, whereas molecular mechanics and molecular orbital geometry optimization procedures minimize energy as a function of many independent atomic coordinates, CONCORD utilizes various relationships between the (internal) atomic coordinates within a given ring and minimizes a composite, univariate strain function. The overall algorithm is quite fast; Rusinko et al.[17] reported the conversion of a corporate database of ≅250000 compounds at the rate of 0.5 s per compound on a VAX 8700, and of course, faster speeds are being achieved on faster machines. The structures generated by CONCORD are usually in very good agreement with the lowest energy structures optimized using molecular mechanics or molecular orbital methods. However, some limitations should be noted. CONCORD will not generate inorganic or metallo-organic structures. Also the program was not intended for use on highly flexible structures (e.g., peptides, macrocycles, polymers) for which the preferred conformation is ill-defined.

Formulation of Flexible Queries. The first step in the procedure for developing flexible queries is selecting a representative molecule from a collection of compounds of known biological activity. Then, the pharmacophoric groups are identified and the rest of the selected molecule is deleted. This identification can be based upon a published pharmacophore, or it can be completely hypothetical: an idea to be tested. The pharmacophoric groups are prioritized with respect to their rigidity or flexibility. The atoms of the rigid (or semi-rigid) parts are then fixed, and the flexible parts of the molecule are anchored to the fixed rigid parts. Distance ranges (encompassing the actual distances in the starting compound) between the flexible parts and fixed parts of the molecule are assigned. At this step, we have developed our "first-draft" search query.

The next step involves conducting searches with the query and subsequently refining distance ranges, deciding on the number of anchors, and adjusting the tolerance on the fixed atoms. The refinement process may include adding other 3D constraints, as needed, to enhance selectivity in the hit list with respect to the desired therapeutic activity. This constitutes the "final" (optimized) search query.

The last step involves a final search with the optimized search query on corporate or proprietary databases to retrieve potential leads.

RESULTS AND DISCUSSION

Let us apply the above procedure to a-well studied case: computer-aided design of angiotensin-converting enzyme (ACE) inhibitors.[18] In general, ACE inhibitors contain one or more of the following molecular functionalities: (a) a zinc-binding ligand, (b) a hydrogen-bond donor, and (c) a carboxyl terminal group.[19] In most cases, the hydrogen acceptor (i.e., hydrogen-bond donor) is an amide carbonyl group[20] and the zinc-binding ligand is either a thiol sulfur or a phosphate or carboxylate oxygen.[21] With the above information, we are now ready to follow the step-by-step procedure to develop a flexible query for ACE inhibitors:

1. Select a representative molecule from a collection of compounds of known biological activity.

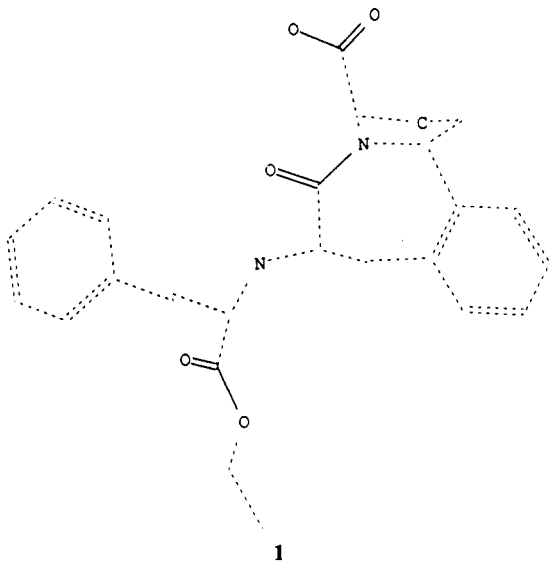
133 of the 16703 3D compounds in MDDR-3D are listed as 'angiotensinase inhibitors'. For this example, we selected one ACE inhibitor, **1** (the pharmacophoric groups are highlighted) (Drug Data Report No. 139250): a drug undergoing biological testing, and patented by Merrell Dow. It is an antihypertensive agent and an in vitro inhibitor of ACE activity.

2. Identify the pharmacophoric groups and delete the rest of the selected molecule.

The terminal carboxyl group and the amide group are clearly identifiable. The zinc binding site must then be the carboxylate group. Also, the secondary amine may be involved in hydrogen bonding.[21] Deleting everything else leaves us with four isolated substructural fragments (see Figure 4a).

3. Prioritize the groups with respect to their rigidity or flexibility.

Examining **1** reveals that the terminal carboxyl group and the amide group are connected with two five- and seven-



1

membered fused rings. Since these two groups are also present in almost all known ACE inhibitors, we can hypothesize (for the purpose of our query building) that they also constitute the rigid section of the molecule. Conversely, the secondary amine and the carboxylate groups lie along a linear chain; therefore, they constitute the flexible section of the molecule.

4. Fix the positions of the rigid parts.

We fix the relative 3D positions of the rigid fragments of **1** as shown in Figure 4b. We fix only the central carbon of the terminal carboxyl group to allow for free rotation of the carboxyl group. We also fix the amide carbonyl group even though the carbonyl carbon is not directly involved in binding to the ACE. The amide carbon can provide another 'hinge' point to be used later in query building. (It is a good strategy to get as many rigid fixed points as possible, since the number of anchors can be an important part of flexible queries in most cases.) At this time, we can verify our hypothesis regarding the flexible and rigid sections of the molecule by running two sets of searches: one with all fragments fixed (except for the oxygen atoms of the two carboxyl groups for reasons stated above) and another with only rigid fragments fixed [i.e., the nonfixed fragments are allowed to be anywhere within the molecule without any restraint (see Figure 4b)]. The tolerance on the fixed points can be adjusted independently via maximum allowed deviation and/or RMS deviation (both in angstroms). In this experiment we kept tolerances on both constraints the same.

In MDDR-3D, 977 compounds contain the substructural components shown in Figure 4a; 78 of these compounds are listed in MDDR-3D as possessing the activity of ACE inhibition. Using this subset of the database, Table I lists the results of searches using the all-fixed query versus rigid-fixed query (see Figure 4b) with varying tolerances. At 0.2-Å tolerance, only one compound is retrieved when all fragments were fixed (the starting compound itself). However, 376 compounds are retrieved when only the rigid fragments are fixed; 61 of these are ACE inhibitors. At 0.7-Å tolerance, 75 of the 78 ACE inhibitors are included in the hit list. These results illustrate that our assignment of rigid and flexible regions for the query is reasonable.

5. Anchor the flexible parts of the molecule to the fixed rigid parts.

By "anchor", we mean connecting a flexible point to a fixed one with a distance constraint. This is done in order to use the fixed atoms as hinge points to define an allowed 3D space for the flexible atoms. Varying the number of anchors and the distance ranges provides some control over the allowed 3D space for the flexible atoms.

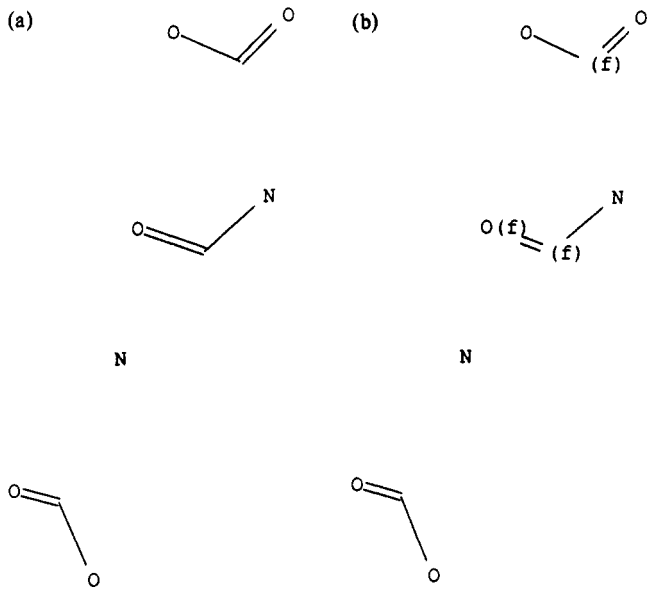


Figure 4. (a) Fragments of **1** remaining after deleting the undesired groups; (b) fixed atoms [denoted by (f)] of fragments that are assumed to belong to the rigid section of the pharmacophore.

Table I. Search Results Using Fixed Atomic Positions with Varying Tolerance

tolerance (Å)[a]		MDDR-3D					
		all atoms fixed			only rigid atoms fixed[b]		
max	RMS	ACE[c]	all[c]	ACE:all, %	ACE[c]	all[c]	ACE:all, %
0.1	0.1	1	1	100	2	6	33
0.2	0.2	1	1	100	61	376	16
0.3	0.3	1	1	100	66	431	15
0.5	0.5	3	3	100	72	680	11
0.7	0.7	3	3	100	75	794	9

[a] Max = maximum allowed distance error for each fixed atom; RMS = maximum allowed root-mean-square deviation for all fixed atoms. [b] Query is shown in Figure 4b. [c] In MDDR-3D (release 90.2), 977 compounds contain all of the substructural fragments shown in Figure 4a; 78 of them are listed as ACE inhibitors.

For example, a single anchor defines a spherical shell centered around the fixed atom (Figure 5a); the radius and the thickness of the shell is defined by the range of distances given for the anchor. Two anchors define a doughnut-like region (*torus*) (Figure 5b); again, the radius and the thickness of the torus is defined by the given distance ranges for the two anchors. Three anchors define two irregularly shaped regions symmetrically placed above and below the plane consisting of the three fixed atoms (Figure 5c). Finally, four anchors define a single irregularly shaped region (Figure 5d).

The distance ranges for the anchors are given by selecting a ± deviation from the actual distances of the fragments from each other. In this example, we used a 4-Å range for all distance constraints (i.e., ≈±2 Å of the actual distance in the molecule).

6. Conduct searches with the query and subsequently refine distance ranges, number of anchors, and tolerance on the fixes, and include other 3D constraints as needed to enhance selectivity in the hit list with respect to the desired therapeutic activity.

The purpose of this refinement step is to increase the ratio of the ACE inhibitors/all-hits (ACE:all) to as large value as possible and *also* with as many ACE inhibitors in the hit list as possible. The optimized query will be very selective toward the compounds with ACE inhibitory activities. This query can then be used to search compounds of unknown activity, and the hit list may provide high quality leads. One can, optionally, relax the tolerance on the fixed atoms in a controlled fashion

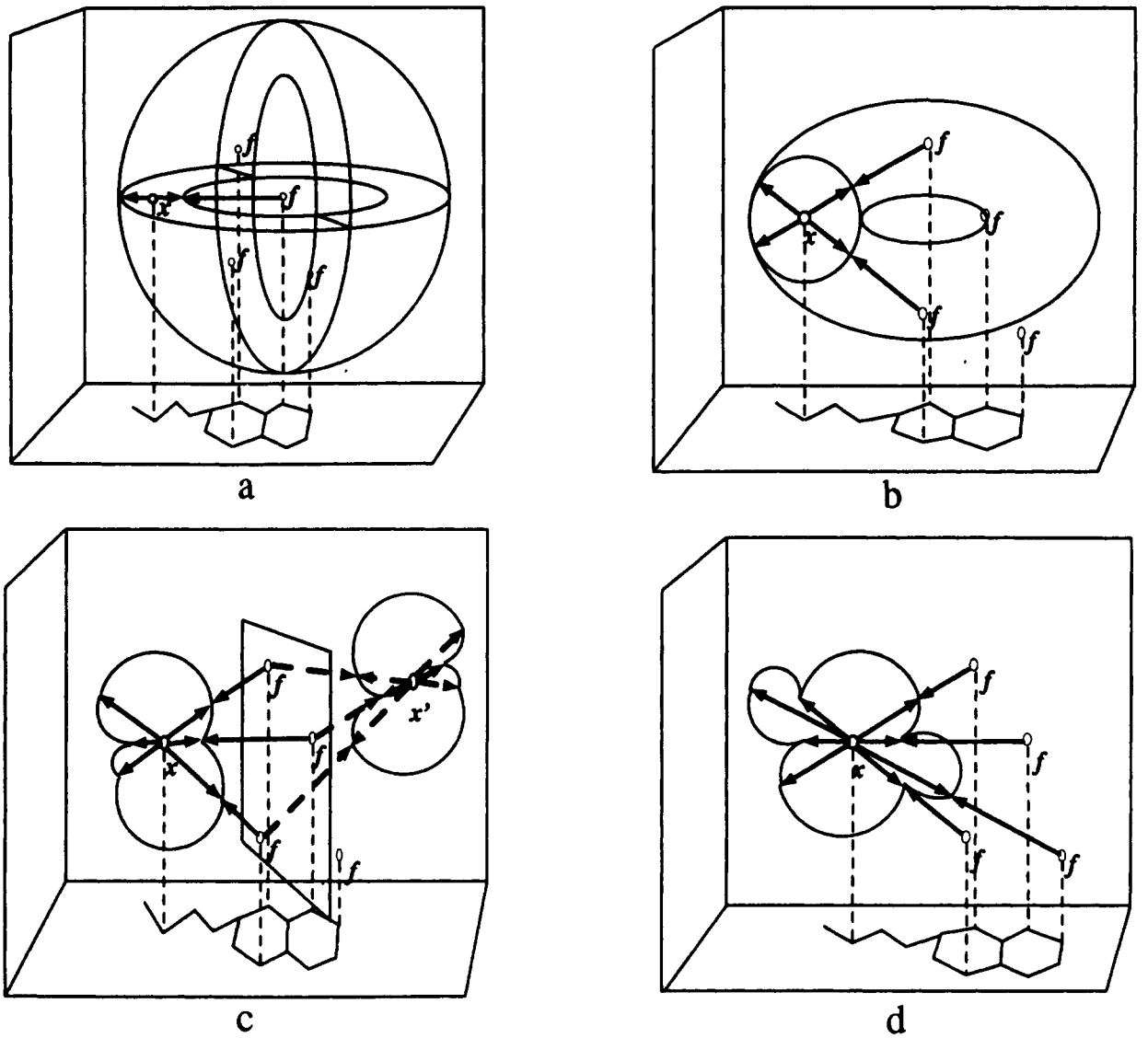


Figure 5. Several examples showing how to anchor a flexible atom to fixed positions: (a) a spherical shell formed by a single anchor; (b) a *torus* formed by two anchors; (c) two irregularly shaped regions (symmetrical around the plane of the three fixed atoms) formed by three anchors; (d) a single irregularly shaped region defined by four anchors.

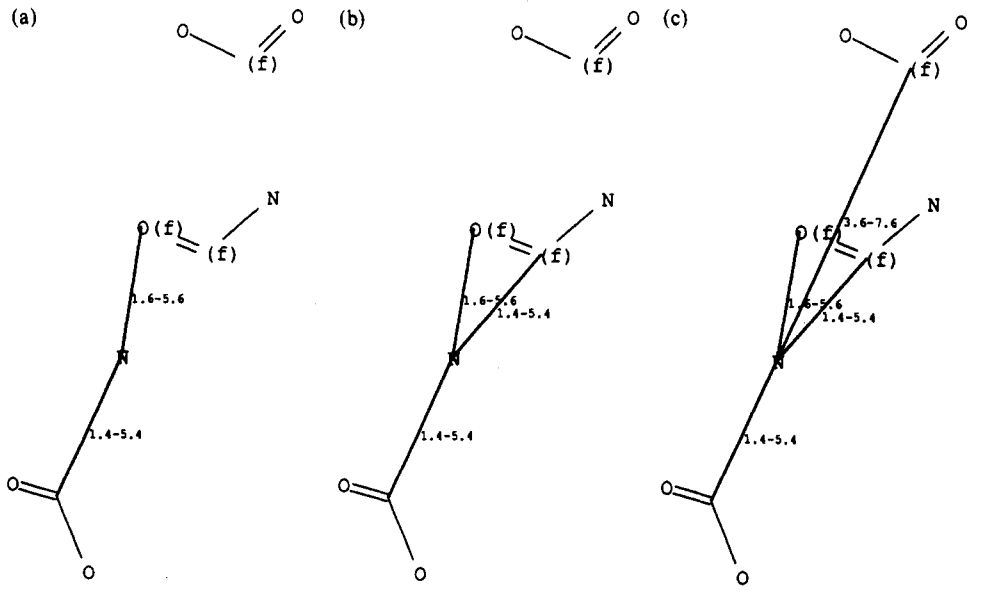


Figure 6. The initial series of flexible queries in an attempt to determine the optimum number of anchors needed for the ACE-inhibitors case: (a) with a single anchor to the central nitrogen (Q-1); (b) with two anchors to the central nitrogen (Q-2); (c) with three anchors to the central nitrogen (Q-3).

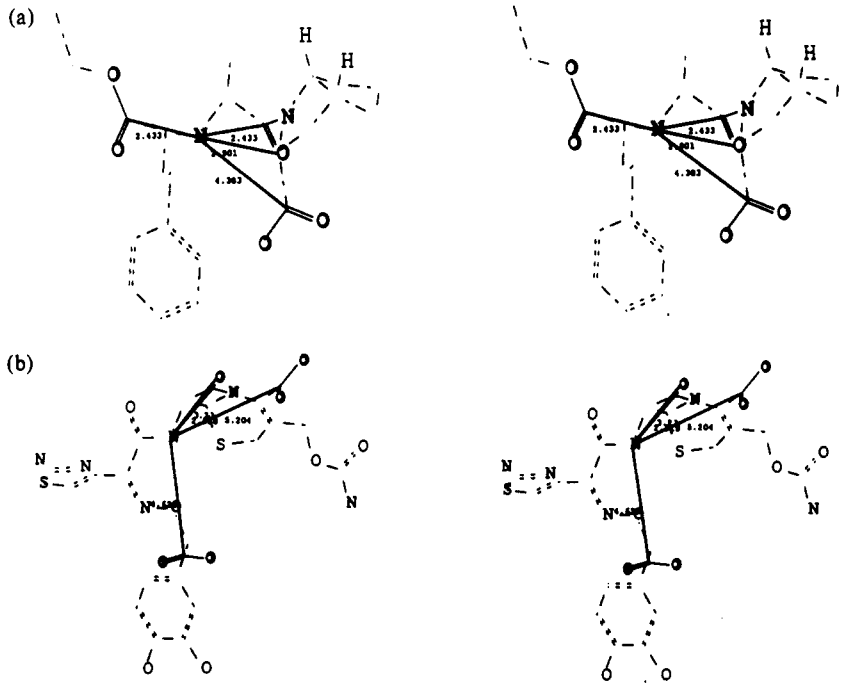


Figure 7. Stereoview of two compounds from the hit list obtained by Q-2; the query objects are overlaid on the structures to show how well they satisfy the constraints: (a) an ACE-inhibitor; and (b) a non-ACE-inhibitor.

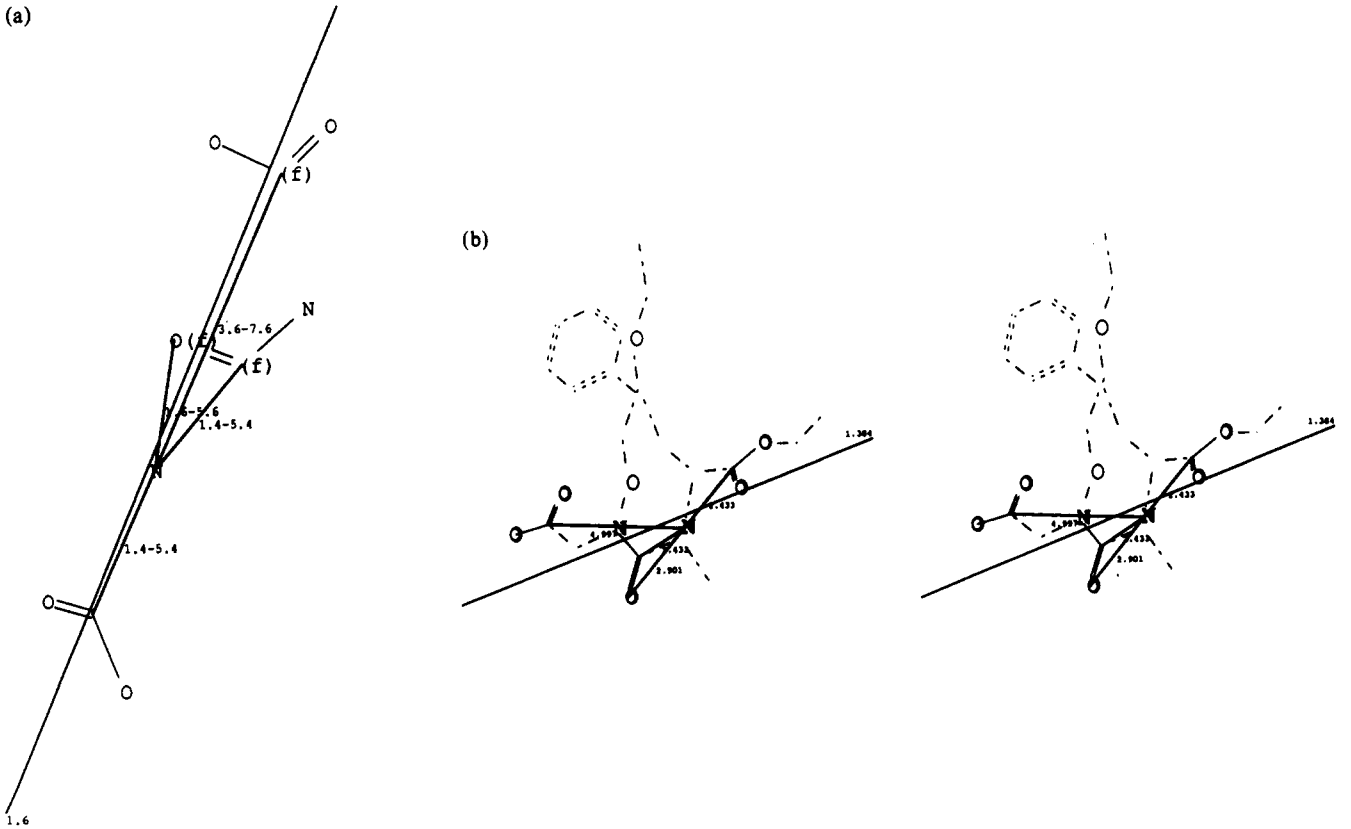


Figure 8. (a) Final flexible query (Q-4), including a linearity constraint for the pharmacophoric groups. (b) Stereoview of an ACE-inhibitor located by the query.

to get more potential leads into the hit list.

Table I shows a dramatic change in the hit list distribution when the tolerance is changed from 0.1 Å to 0.2 Å. The number of compounds in the hit list increased from 6 to 376; also, the number of ACE inhibitors increased from 2 to 61. Although ACE:all is relatively high ($2 \times 100/6 = 33\%$) at the 0.1-Å tolerance level, the number of compounds in the hit list is too few.

Therefore, we will refine the query at the tolerance set of 0.2 Å to improve the ACE:all ratio and to keep as many ACE inhibitors as possible in the hit list. We test three queries (Figure 6a–c) with different numbers of anchors to the central nitrogen (Table II). While the number of ACE inhibitors in the hit list remains at 60 (Q-1), the total number of hits also remains the same at 204 when we use two anchors to the central nitrogen (Q-2) and decreases to 181 when we use three anchors (Q-3). Evaluating the ACE inhibitors in the hit list and comparing them with the non-ACE inhibitors suggests

Table II. Search Results Using Flexible Queries and Comparison with Pharmacophoric Searches

		MDDR-3D		
query	tolerance,[a] (Å)	ACE	all	ACE:all, %
Q-1	0.2	60	204	29
Q-2	0.2	60	204	29
Q-3	0.2	60	181	33
Q-4	0.2	60	83	72
Q-4	0.3	64	91	70
Q-4	0.5	64	93	69
Q-4	0.7	64	102	63
Q-4	1.0	64	114	56
Q-4	2.0	64	120	53
Q-5	n/a	72	645	11
Q-6	n/a	58	560	10
Q-7	n/a	24	165	15

[a] Tolerance indicates the value used for both; max = maximum allowed distance for each fixed atom; RMS = maximum allowed root-mean-square deviation for all fixed atoms from the corresponding fixed positions; and n/a = not applicable.

a way to further refine our query. Figure 7a displays an ACE inhibitor overlaid with the triply anchored query (Q-2), and Figure 7b displays one of the non-ACE inhibitors overlaid with the query. Note that the carboxylate group on the flexible region of the non-ACE inhibitor is incorrectly positioned with respect to the semi-rigid section. To eliminate such spurious hits, we can refine our query by adding an angle or dihedral angle constraint. In this case we add a linearity constraint. We position carboxyl carbons, the amide carbonyl oxygen, and the secondary amine nitrogen on a straight line. By giving a 1.6-Å RMS tolerance we define an inclusion volume (resembling a crude cylinder) for the selected groups (Q-4, Figure 8a). This query hits all of the 60 ACE inhibitors of Q-3 (see Figure 8b) while reducing the total number of hits from 181 to 83 (Table II). Hence Q-4 at the 0.2-Å tolerance level represents the optimized query for this case.

We accomplished our first objective by increasing the ACE:all to 72% while keeping a relatively large number of known ACE inhibitors (60) in our hit list. However, as our primary goal was to generate good leads, we will relax our grip at the rigid section of the query (Q-4) by gradually increasing the tolerance on the fixed positions. Table II lists the search results. As we increase the tolerance, the number of known ACE inhibitors stays at 64 while the total number of hits gradually increases.

7. Perform a final search with the refined query on corporate or proprietary databases to retrieve potential leads.

These results illustrate that useful 3D searches can be performed with flexible queries that have tighter restrictions on rigid portions of queries and looser restrictions on flexible portions. Such queries can retrieve a high percentage of the active compounds in a diverse database. Although additional structures are retrieved with looser tolerances, these additional structures may be prime candidates for testing in a screen for ACE inhibitors.

Finally, we will compare the search results of our flexible queries with a typical pharmacophoric search (which does not take into account the flexibility of the compounds in a static database). Mayer et al. proposed a pharmacophore for ACE inhibitors using four-point distances from the zinc atom in ACE.[22] Figure 9a shows the MACCS-3D interpretation of the pharmacophore. We indicate the location of the zinc by a point that is 4.5 Å away from, and along the direction of, any two atoms in the molecule (the 4.5-Å distance was obtained by optimizing the query to yield the largest number of ACE inhibitors). We place a small exclusion sphere (1-Å radius) around the dummy point representing the zinc atom.

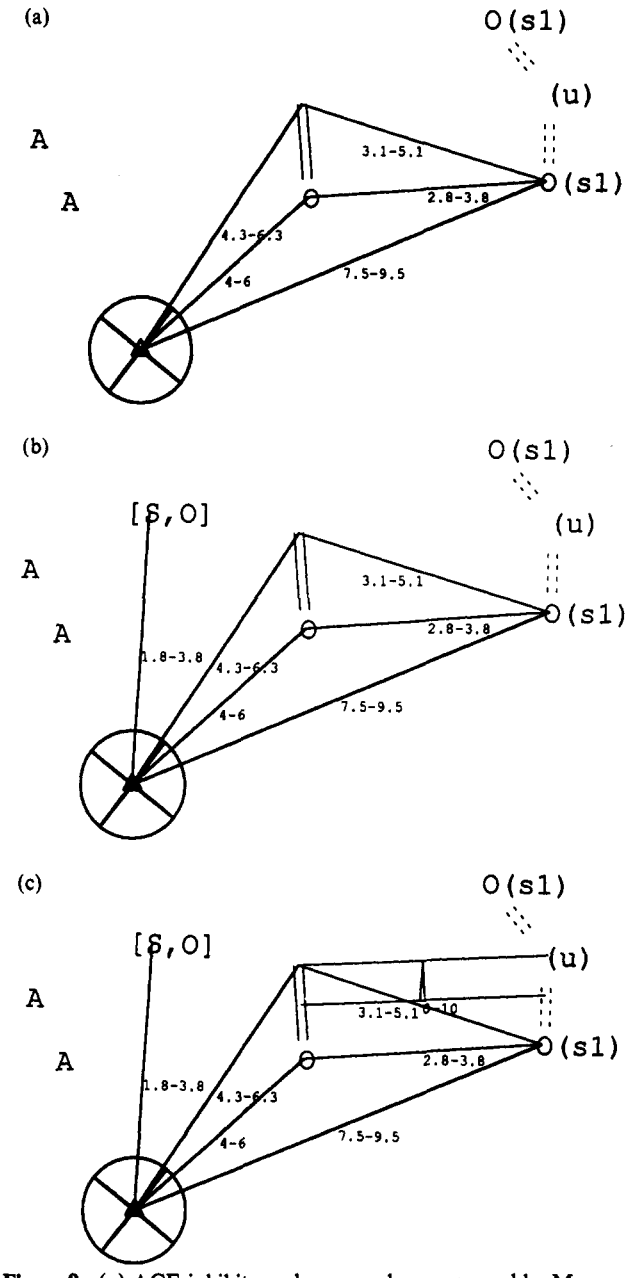


Figure 9. (a) ACE-inhibitors pharmacophore proposed by Mayer et al.[22] (Q-5); (b) modified query, including the zinc-binding site (Q-6); (c) query further modified with an additional torsional constraint (Q-7). Position of the zinc atom is indicated by a dummy point at the center of the exclusion sphere.

We use the distances proposed by Mayer et al.[22] in the query with ≈±1-Å range (see Figure 9a). Although this query (Q-5) hits slightly more ACE inhibitors (72) than our previous query (Q-4), the ACE:all is extremely low (11%) leaving the scientist with many spurious hits which are costly and time-consuming to pursue. An attempt to improve the selectivity by including the zinc binding atoms into the query—at about 2.8 Å from the zinc atom as proposed by Marshall[23]—(see Figure 9b) decreases the number of ACE inhibitors as well as the total number of hits in almost the same ratio. The resulting query (Q-6) has about the same ACE:all (10%). A further attempt to improve the selectivity of this query is done by incorporating a torsional constraint to keep the amide carbonyl and the terminal carboxylate groups in gauche conformation as suggested by Mayer[22] (Figure 9c). The final query Q-7 improved the selectivity slightly (15% ACE:all) but it also severely cut the number of ACE inhibitors in the list to 24. Figure 10 displays **1** with Q-7 mapped onto it. In all three searches,

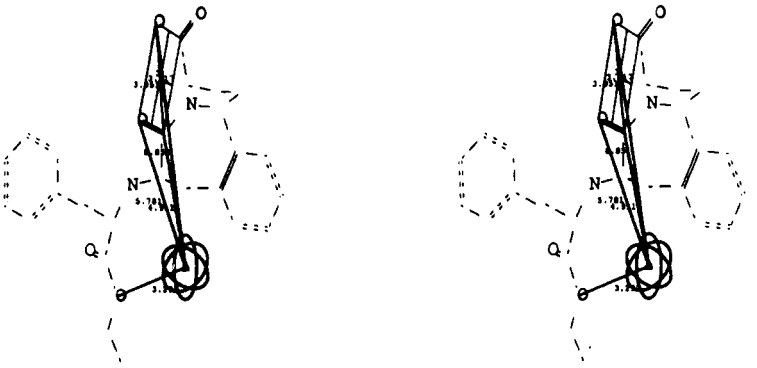


Figure 10. Stereoview of the starting ACE-inhibitor hit by the query Q-7.

however, the selectivity is considerably less than our optimized flexible query (Q-4, ACE:all = 72%).

While demonstrating the utility of flexible queries, we do not suggest abandoning pharmacophoric searches; the two techniques are complementary to each other and should be used together. Because the 2D substructural fragments used in the two approaches are slightly different, the hit lists may contain somewhat different compounds. In fact, of the 58 ACE inhibitors obtained from Q-6, 16 are not found among the 60 ACE inhibitors of the list from Q-4. Similarly 18 compounds in the list from Q-4 are not found within the list from Q-6.

CONCLUSIONS

Our goal was to demonstrate the use of flexible queries as a viable and easily implemented approach to addressing the conformational flexibility issue in 3D searching. There are certain assumptions, limitations, and advantages which are implicit in this approach. First, it is important to realize that we are developing a pharmacophoric query and not necessarily a biologically relevant pharmacophore. The main goal of the analysis is to find a 3D query which will successfully screen a high proportion of active compounds in a given database.

Another assumption is that fixed and flexible regions exist in the pharmacophore. This is generally true for all but the simplest pharmacophores. Even for a simple three-point pharmacophore, it is likely that one of the points will be more uncertain or flexible in its position than the others. As the pharmacophore becomes more complex and as interatomic distances increase, the assumption becomes increasingly valid. Clearly, it is important to use compounds which have both rigid and flexible regions in developing the pharmacophore model. The justification for this dates back to the introduction of the active analogue approach of Marshall.[24]

Keeping these assumptions in mind, we have shown that the flexible query approach can be successfully used to model a pharmacophore query which can then be applied to a commercial database of 3D conformations to yield a large proportion of lead compounds. We have presented a stepwise procedure for developing such a query. It should be noted that certain steps in this procedure lend themselves to automation. Particularly, in MACCS-3D, it is possible to write command sequences which can systematically set distances, other constraints, and even query atoms. It can then run a search and, based on the results, vary the constraints and atoms, eventually arriving at an optimal query. This automation may not be desirable in the initial steps of query generation, since it does not allow the chemist any influence over the query, but for later refinement steps, it could be quite useful.

In many ways, 3D searching is in its infancy. As flexible queries are eventually replaced by flexible searching, one can expect to see parallel developments in search-time property estimation, greater use of binding and transport energetics as a part of the search process, 3D similarity and molecular shape calculations, and even search-time modification of structures to suggest new leads.[2,13a]

ACKNOWLEDGMENT

We thank Thomas E. Moock, who originally suggested the idea of "fixing" the rigid atoms of a molecule. These fixes became an important part of the "flexible queries" described in this paper.

Registry No. ACE, 9015-82-1.

REFERENCES AND NOTES

(1) Martin, Y. C.; Bures, M. G.; Willett, P. Searching Databases of Three-Dimensional Structures. In *Reviews in Computational Chemistry*; Boyd, D., Lipkowitz, K., Eds.; VCH Publishers: New York, 1990; Chapter 6, pp 213–263.

(2) Pearlman, R. S. 3D Searching: An Overview of a New Technique for Computer Assisted Molecular Design. In *Emerging Technologies and New Directions in Drug Abuse Research*; Rapaka, R., Ed.; Row Scientific: Washington, DC, 1991; pp 62–77.

(3) Sheridan, R. P.; Rusinko, A.; Nilakantan, R.; Venkataraghavan, R. Searching for Pharmacophores in Large Coordinate Data Bases and Its Use in Drug Design. *Proc. Natl. Acad. Sci. U.S.A.* **1989**, *86*, 8165–8169.

(4) Sheridan, R. P.; Nilakantan, R.; Rusinko, A.; Bauman, N.; Haraki, K. S.; Venkataraghavan, R. 3DSEARCH: A System for Three-Dimensional Substructure Searching. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 255–260.

(5) Van Drie, J. H.; Weininger, D.; Martin, Y. C. ALADDIN: An Integrated Tool for Computer-Assisted Molecular Design and Pharmacophore Recognition. *J. Comput.-Aided Mol. Des.* **1989**, *3*, 225–251.

(6) Bures, M. G.; Black-Schaefer, C.; Gardner, G. The Use of Molecular Modeling Techniques To Discover Novel Auxin Transport Inhibitors. *Proc. Natl. Acad. Sci. U.S.A.*, in press.

(7) (a) Christie, B. D.; Henry, D. R.; Güner, O. F.; Moock, T. E. MACCS-3D: A Tool for Three-Dimensional Drug Design. In *Online Information 90*; 14th International Online Information Meeting Proceedings; Raitt, D. I., Ed.; Learned Information, Oxford, 1990; pp 137–161. (b) Moock, T. E.; Christie, B. D.; Henry, D. R. MACCS-3D: A New Database System for Three-Dimensional Molecular Models. In *Chemical Information Systems*; Bawden, D., Mitchell, E. M., Eds.; Ellis Horwood: Chichester, 1990; pp 42–49. (c) Güner, O. F.; Dumont, L. M. 3D Searching in Computer-Aided Drug Design. In *Pharmaceutical Manufacturing International 1991*; Barber, M. S., Barnacal, P. A., Eds.; Sterling Publications: London, 1990; pp 65–68.

(8) (a) MDDR-3D (version 90.1); available from Molecular Design Limited, San Leandro, CA. (b) Grethe, G.; Dumont, L. M. A New Electronic Database. *Drug News Persp.* **1989**, *2*, 488.

(9) Pearlman, R. S. Rapid Generation of High Quality Approximate 3D Molecular Structures. *Chem. Des. Auto. News* **1987**, *2*, 1–7.

(10) Pearlman, R. S.; Rusinko, A.; Skell, J. M.; Balducci, R. CONCORD; Tripos Associates Inc.: St. Louis, MO, 1987.

(11) Ahrens, E. K. F. Customization for Chemical Database Applications. In *Chemical Structures*; Warr, W. A., Ed.; Springer-Verlag: Berlin, 1988; pp 97–111.

(12) Güner, O. F.; Hughes, D. W.; Dumont, L. M. An Integrated Approach to Three-Dimensional Information Management with MACCS-3D. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 408–414.

(13) (a) Martin, Y. C. Computer Design of Potentially Bioactive Molecules by Geometric Searching with ALADDIN. *Tetrahedron Comput. Methodol.* **1990**, *3*, 15–25. (b) Martin, Y. C. ALADDIN: A Real Tool for Structure-Based Drug Design. In *Crystallographic and Modeling Methods in Molecular Design*; Bugg, C. E., Ealick, S. E., Eds.; Springer

Verlag: New York, 1990; pp 254-263. (c) Martin, Y. C. Beyond Graphics: ALADDIN, A Computer Tool for Drug Design. In *Frontiers in Drug Research*; Jensen, B., Jorgensen, F. S., Kofod, H. Eds.; Alfred Benzon Symposium 28; Munksgaard: Copenhagen, 1990; pp 222-229.
(14) (a) Murrall, N. W.; Davies, E. K. Conformational Freedom in 3-D Databases. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 312-316. (b) Davies, E. K.; Upton, R. M. In *Online Information 90*, 14th International Online Information Meeting Proceedings; Raitt, D. I., Ed.; Learned Information: Oxford, 1990; p 129.
(15) Prous, J. R. *Drug Data Report*, Vol. 11; Science Publications: Barcelona, 1989.
(16) Allen, F. H.; Kennard, O.; Watson, D. G.; Brammer, L.; Orpen, A. G.; Taylor, R. Tables and Bond Lengths Determined by X-ray and Neutron Diffraction. *J. Chem. Soc. Perkin Trans. 2* **1987**, S1-S19.
(17) Rusinko, A.; Sheridan, R. P.; Nilakantan, R.; Haraki, K. S.; Bauman, N.; Venkataraghavan, R. Using CONCORD To Construct a Large Database of Three-Dimensional Coordinates from Connection Tables. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 251-255.
(18) Hangauer, D. G. In *Computer-Aided Drug Design, Methods and Applications*; Perun, T. J., Propst, C. L., Eds.; Marcel Dekker, Inc.: New York, 1989; pp 253-295.
(19) Saunders, M. R.; Tute, M. S.; Webb, G. A. A Theoretical Study of Angiotensin-Converting Enzyme Inhibitors. *J. Comput.-Aided Mol. Des.* **1987**, *1*, 133-142.
(20) Andrews, P. R.; Carson, J. M.; Caselli, A.; Spark, M. J.; Woods, R. Conformational Analysis and Active Site Modeling of Angiotensin-Converting Enzyme Inhibitors. *J. Med. Chem.* **1985**, *28*, 393-399.
(21) Petrillo, E. W., Jr.; Ondetti, M. A. Angiotensin-Converting Enzyme Inhibitors: Medicinal Chemistry and Biological Actions. *Med. Res. Rev.* **1982**, *2*, 1-41.
(22) Mayer, D.; Naylor, C. B.; Motoc, I.; Marshall, G. R. A Unique Geometry of the Active Site of Angiotensin-Converting Enzyme Consistent with Structure-Activity Studies. *J. Comput.-Aided Mol. Des.* **1987**, *1*, 3-16.
(23) Marshall, G. R.; Motoc, I. In *Molecular Graphics and Drug Design*; Burgen, A. S. V., Roberts, G. C. K., Tute, M. S., Eds.; Elsevier Science Publishers: Amsterdam, 1986; p 115.
(24) Marshall, G. R.; Barry, C. D.; Bosshard, H. E.; Dammkoehler, R. A.; Dunn, D. A. In *Computer Associated Drug Design*; ACS Symposium Series 112; American Chemical Society, Washington, DC, 1979; p 205.
(25) Lloyd, E. J.; Andrews, P. R. A Common Structural Model for Central Nervous System Drugs and Their Receptors. *J. Med. Chem.* **1986**, *29*, 453.

Computer-Assisted Study of the Relationship between Molecular Structure and Surface Tension of Organic Compounds

DAVID T. STANTON† and PETER C. JURS*
Chemistry Department, 152 Davey Laboratory, Penn State University, University Park, Pennsylvania 16802

Received July 15, 1991

Computer-assisted methods are applied to the study of the relationship between molecular structure and observed surface tension of small organic alkanes, alkyl esters, and alkyl alcohols. Features of these molecules are encoded using a wide variety of topologic, geometric, and electronic descriptors. The simple correlation between these descriptors and observed surface tension values is examined to gain insight as to which molecular features most influence the observed surface tension. Multivariate linear regression models for each functional group class are also examined. The results of the examination of both the simple correlations and the regression models suggest that molecular surface area is an important feature. The results also show that many descriptors provide surface area information which is specific to particular portions of the molecule, and that this information provides better results in modeling surface tension than the van der Waals or solvent-accessible surface area. Finally, a multiple linear regression model is developed for a combined set of alkanes, alkyl esters, and alkyl alcohols which yields good results for predicting the surface tension of similar compounds.

INTRODUCTION

Surface tension, like normal boiling point or chromatographic retention, is a physical property which is a function of molecular structure. However, very little has been published in the chemical literature concerning the relationship between the structure of a molecule and the observed surface tension at a given temperature. The purpose of the research described here was twofold. The first goal was to establish that surface tension is a physical property that can be studied in the same fasion as other properties using quantitative structure-property relationship (QSPR) techniques and the tools available in the ADAPT software system.[1,2] The second goal of this work was to determine if such a study could shed light on the structure-property relationship involving surface tension.

The reasons for studying the relationship between molecular structure and a physical property such as surface tension are very similar to the reasons for studying normal boiling points or any other property. The material of interest may be in short supply or the experimental procedure itself may be too time consuming or expensive to be performed for more than just a few compounds. However, with the aid of a carefully developed predictive model, the values for the property of interest can be quickly and accurately estimated. But among the usual advantages which can be obtained from the modeling of a given physical property, the ability to learn something about how the structural features of a molecule can affect that property is possibly the most important. This is especially true in the case of surface tension. There have been many generalized statements made which associate polar and hydrogen-bonding intermolecular interactions with increased surface tension of pure liquids.[3,4] Surface tension has also been noted to increase as a function of molecular weight for a set of congeners.[5] However, very little more has been published in the chemical literature concerning the structure-property relationship for surface tension of pure organic compounds. Therefore, it was of interest to employ QSPR techniques to expand our understanding in this area.

In a previous paper, a brief study of the relationship between molecular structure and observed surface tensions for a small and diverse set of organic compounds was reported.[6] The results of that study suggested that it was possible to model surface tension. However, the results also suggested that additional work was necessary to improve the accuracy of the resulting models. Part of the problem associated with the accuracy of the model developed in that study involved the small size and wide diversity of the compounds in the dataset involved. In order to study the structure-property relationship for surface tension effectively, it was necessary to select a dataset of reasonable size and minimum diversity. In this way

† Present address: Norwich Eaton Pharmaceuticals, Inc., P.O. Box 191, Norwich, NY 13815.

0095-2338/92/1632-0109$03.00/0