

J-CAMD 373

## Strategies for the determination of pharmacophoric 3D database queries

John H. Van Drie

*Pharmacia & Upjohn Inc., Kalamazoo, MI 49001, U.S.A.*

Received 10 May 1996

Accepted 24 July 1996

*Keywords:* 3D database; Pharmacophore; Database query; Selectivity principle; D<sub>2</sub>-antagonist pharmacophore; ACE pharmacophore; RGD pharmacophore; β<sub>2</sub>-antagonist pharmacophore

---

### Summary

Strategies are described for constructing pharmacophoric 3D database queries, based on a series of active and inactive analogs. The results are highly selective database queries, which are consistent with the generally accepted pharmacophore for a number of systems. The foundation of these strategies is the method of Mayer, Naylor, Motoc and Marshall [J. Comput.-Aided Mol. Design, 1 (1987) 3] for inferring a unique binding geometry for angiotensin-converting enzyme (ACE) inhibitors. The strategies described here generalize their approach to cases where the chemical features responsible for binding are not a priori apparent, and to cases where the binding geometry deduced by that method is not unique. The key new insight, the selectivity principle, is to rank the multiple solutions produced by the method of Mayer et al. by their selectivity, a value that is related to the proportion of a database that is returned as a database hit list. Retrospective analyses are described for D<sub>2</sub>-antagonists, ACE inhibitors, fibrinogen antagonists, and β<sub>2</sub>-antagonists.

---

### Introduction

Searching 3D databases to discover novel activities of existing compounds has been widely applied in recent years. Two distinctly different approaches are being used: (i) shape-based methods, such as DOCK [1], in which a protein structure is used to formulate the database query, to search for compounds whose structure complements the receptor's steric characteristics; and (ii) pharmacophore-based methods, which search for compounds whose structure satisfies a certain pharmacophoric pattern, i.e. a specification of the geometric arrangement of a set of constraints formulated on a set of functional groups [2–9]. In methods of the second type, one seldom sees detailed discussions of the problem of determining the pharmacophore, i.e. constructing an appropriate query of such a geometric arrangement of a set of functional-group constraints. Generally, what is described is taking pharmacophores from the literature and using them as 3D search queries, or taking them from a protein crystal structure. The purpose of this paper is to highlight the problems associated with determining pharmacophoric search queries for 3D database searches of the second type *for those cases where no protein structure information is available*, and to describe strategies for dealing with these problems.

This problem has been the focus of attention of a number of investigators. The strategies described here have been developed over many years of experience, and have been successfully applied to a wide range of problems, both retrospectively and prospectively; some retrospective analyses will be described here. In particular, 'flexible 3D database search systems' (ones which take into account the flexibility of each candidate molecule in the database) allow one to apply these strategies to discover pharmacophores from a set of active analogs, along the lines of traditional active-analog receptor mapping [10]. This problem has received considerable attention recently (reviewed, for example, in Ref. 11), and is frequently referred to as the 'pharmacophore identification problem' or the 'pharmacophore recognition problem'. What is presented here is less automated than what is usually attempted under that rubric, but the intent is that the strategies presented here will assist in solving the pharmacophore recognition problem.

### Overview of the problems of constructing pharmacophoric 3D database queries

Generally, one faces four problems in composing a 3D database query:

(1) Which functional groups should one choose, i.e. what set of topological constraints?

(2) Which type of geometric relationships should one impose, e.g. interatomic distances, angular relationships between two pendant substituents, the angle of a lone pair to the plane of a phenyl ring?

(3) What average distance should separate the features, or by what average angle should the features be arranged?

(4) What tolerances should one impose on these geometric relationships?

The first problem is often dealt with in a desultory fashion, and often relies heavily on SAR (structure–activity relationships) developed long after the most promising compounds have been synthesized. For example, in the treatment of ACE inhibitors by Mayer et al. [12] they chose the sulfide, the proline carboxyl, and the carbonyl of the amide bond adjacent to the proline. The selection of such features is easily made, given the decade or more of SAR that have accumulated after the discovery of the first potent ACE inhibitor, captopril [13], but such a selection would have been difficult to make given the peptidic inhibitors known prior to captopril. Similarly, the 3D database query for D<sub>2</sub>-antagonists that we originally employed [4] was based on the extensive SAR analyzed by Seeman et al. [14]. Since the greatest potential application of 3D database searching is in identifying new leads early in drug development projects, the best strategies for selecting features are those that do not require vast amounts of SAR. Regarding the problem of selecting geometric relationships, some approaches to 3D database searching offer only interatomic or -feature distance constraints [2,3,5] or only angular relationships [15]; this simplifies the problem by reducing one's options, but the more general approaches allow one to specify a combination of distances to the centers of features, three-center angles and torsion angles, and angular relationships to least-squares planes. A review of many of the published pharmacophores and receptor maps makes it clear that all such relationships are useful in describing the ligand–receptor recognition; hence, this is a problem best confronted and not ignored. This is a more subtle problem than one might expect a priori.

A simple approach for dealing with the last two problems, establishing average values for interfeature distances, etc., and establishing their tolerances, which has been advocated since these 3D database search techniques were introduced [16], is to establish the tightest geometric relationships which still allow the query to hit all the known active molecules. One performs this by building a small database of actives, and iteratively refining the query until the tolerances cannot be made tighter without losing some of the actives as hits. The primary difficulty with this approach is that the resulting queries are not especially discriminating, e.g. with this method it is difficult to compose a query for  $\beta_1$ -antagonists that does not also hit

$\beta_1$ -agonists. Also, such queries routinely suffer from the ‘multiple-mapping problem’, in which any given active may ‘map’ to the query in many ways, thus obscuring which groups on the active molecule are contributing to binding (by ‘map’, we mean the assignment of features in the query to features on the molecule). Ideally, queries should map to all active molecules in one, unique way, to allow unambiguous judgements on how to synthetically modify any hit from a database search to improve the activity. (Of course, biologically it is possible for a molecule to bind to its receptor in multiple ways. In such cases one would expect multiple mappings. However, this should be the exception, not the rule.) It is not uncommon to compose literature pharmacophores as 3D search queries, and to discover that they map to actives in multiple ways.

An especially intriguing method for determining distance ranges (average interfeature distances and their tolerances) was reported in the literature many years ago by Mayer et al., where they successfully identified a unique binding geometry of ACE inhibitors, by looking for a region of overlap common to all actives in the space of interfeature distance constraints. Their method offers a very refined determination of the distances separating the features, taking fully into account the conformational flexibility of the molecules. Figure 1 depicts schematically how their method works. Suppose we have three features of interest: (A) a carbonyl oxygen, (B) a carboxylate, and (C) a sulfide. Let us consider the space of all possible distances achievable between A and B,  $d_{AB}$ , and the space of all possible distances achievable between B and C,  $d_{BC}$ . We can plot for each molecule a region of space for which this molecule may adopt a conformation with the values of  $d_{AB}$  and  $d_{BC}$ , which we schematically represent as a circle in Fig. 1. Circle 1 represents the conformational space of active analog 1, circle 2 represents that of active analog 2, and circle 3 that of active analog 3. The overlap region of these circles represents the space of possible binding geometries which are common to all the actives, and hence is likely to represent the binding geometry at the receptor. The more active analogs included, the more tightly defined this overlap region becomes. Figure 1 just shows two dimensions,  $d_{AB}$  and  $d_{BC}$ ; this method is applicable to any number of dimensions.

The primary drawback to general applicability of their method is that it is infrequently the case that a *unique* binding geometry arises, as in the ACE inhibitor data set of Mayer et al.; in general, one sees multiple solutions, i.e. multiple common overlap regions, especially when one considers all possible mappings of the features of interest. One is left with the problem of choosing among the many possible solutions. This problem is especially troublesome when one's actives are all highly flexible and feature-rich molecules, such as peptides. Another difficulty in identifying a binding geometry is that, in some cases, a unique

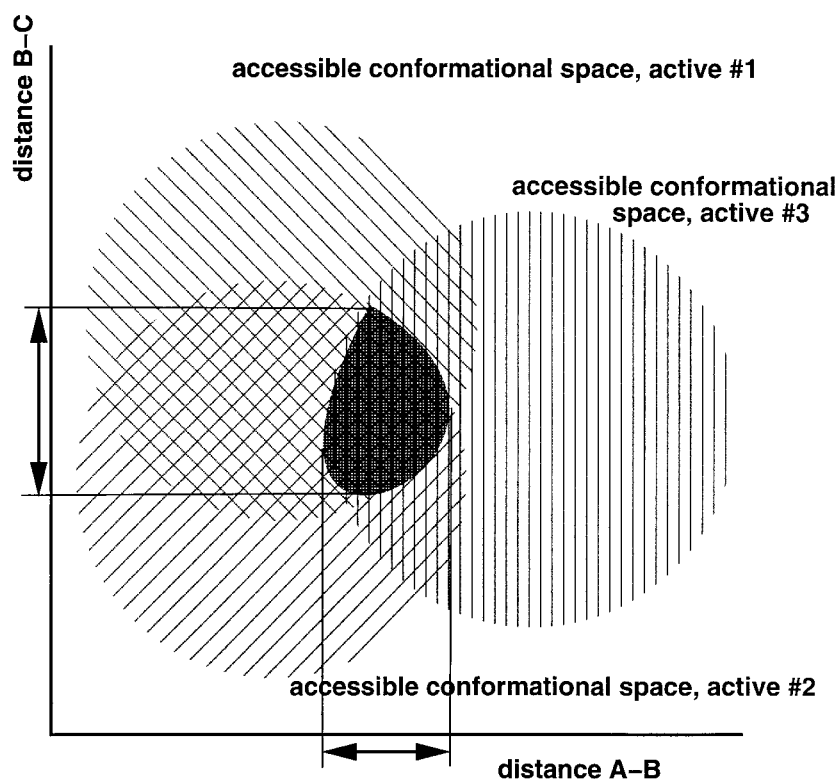


Fig. 1. Schematic depiction of the method of Mayer, Naylor, Motoc and Marshall [12] for determining a geometry common to a set of three molecules. The overlap region defines distance ranges from A to B and from B to C for the pharmacophore.

binding geometry is not correct, and multiple solutions may be required: the molecules may bind in multiple binding modes, or the assay may not yet have been refined, such that multiple receptor subtypes give rise to the binding data one has collected. Ideally, a strategy for constructing 3D database queries would detect when that may be the case, and identify two queries, one which hits one subset of the molecules and another which hits the remainder.

### Strategies for handling some of these problems

The following strategies for dealing with the aforementioned problems in the development of 3D database queries are proposed. Significant chemical judgement is still required in applying these strategies. The starting point is a small (3–50) number of actives. The end result is usually a small number of useful, highly selective 3D database queries, ranked in order of their anticipated utility.

#### Feature identification

Use a standard library of features, based on the physical chemistry of ligand–receptor binding:

- (1) hydrogen-bond donors or acceptors;
- (2) groups capable of forming salt-bridges (groups with positive or negative charges, or capable of acquiring them in solution at physiological pH);

(3) groups capable of forming pi-stacking interactions with the receptor; and

(4) hydrophobic groups.

We will always consider *all possible mappings* of these features to any active molecule. By ‘mapping’, recall that it is an assignment of that feature to atoms of the molecule. For example, the dipeptide Asp-Ala would have hydrogen-bond donors identified on the  $\delta$ -carboxyl, the terminal carboxyl, and the amide bond; hydrogen-bond acceptors identified on those three groups as well; a negative-charge group on either carboxyl; a hydrophobic group on the alanine side-chain; a positive-charge group on the N-terminus; and no pi-stacking mappings. The first two types are ‘oriented’ features (i.e. those which will bind to a receptor in a preferred orientation). The others are nonoriented, or isotropic, features. Only distance constraints may be made to nonoriented features; oriented features can have their orientation relative to one another specified by three-center angle and four-center torsion constraints.

#### Constraint selection

Use distance constraints between the center of each pair of features. Angle and torsion constraints between oriented features are applied in the final stages based on a chemical inspection of the resulting binding orientations.

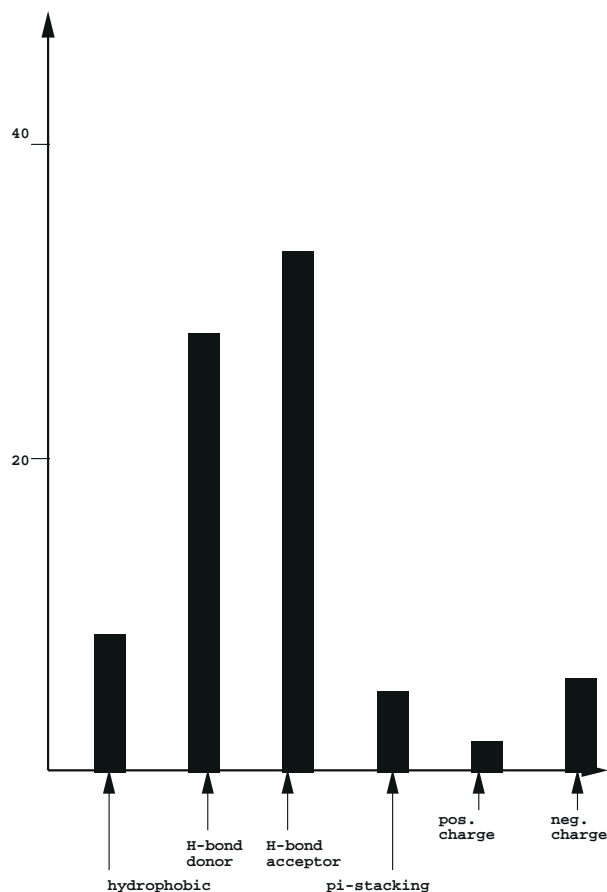


Fig. 2. Tabulation of occurrences of types of ligand-receptor interactions among 10 randomly chosen entries in the Protein DataBank.

#### Identification of possible binding geometries

Apply the method of Mayer et al. to search the multidimensional space of all distances and identify those regions of space which are common to most or all of the actives.

#### Ranking multiple solutions

Use *selectivity* as a ranking criterion. Those queries which are the most selective are preferred. Selectivity is computed by determining the proportion of hits  $q$  that a query returns when applied to a large database composed of molecules structurally like the actives; ranking is based on the selectivity index  $q^N$ , where  $N$  is the number of actives which this pharmacophore hits.

#### Iterative, progressive construction from queries with small numbers of features to large

Construct queries initially from any two features ('dyads'), apply the method of Mayer et al., rank the results, and judge the best queries by their ranking and their ability to map sensibly (i.e. the dyad should map consistently across the actives, it should map in either a unique way or a small number of ways to each active, and each mapping should be chemically reasonable and consistent with any known SAR). From the initial dyads,

consider all possible additions of a third feature ('triads'), and repeat this procedure, *keeping the geometric constraints among the original dyad features intact*. Enlarge the number of features (e.g. extend the triads to tetrads, keeping the constraints of the original triad intact) until no consistent additional feature can be found among most or all of the actives, or until adding a new feature does not significantly improve the selectivity of the query as judged by the selectivity index.

The use of the first strategy may be justified simply by observing the classes of receptor-ligand interactions that one sees experimentally. Figure 2 shows a histogram of the types of receptor-ligand interactions one sees from 10 entries randomly selected from the Protein DataBank. These six types of interactions are seen to dominate. A random number generator was used to generate indexes into an alphabetically sorted list, yielding 30 PDB entries. The first 10 which had a single ligand and were not multimeric and did not contain a prosthetic group were used: 1ASE, 1CNX, 1CRQ, 1DHI, 1ERB, 1MNS, 1MRK, 1NGB, 1NGF, 2IFB. The interactions the ligand made with protein atoms out to 4 Å were graphically inspected. While for our purposes these results are adequate, it should be noted that this procedure yielded two nucleotide-triphosphate ligands, so it may not be representative of the PDB as a whole, especially with regard to the counts of H-bond acceptors and donors. A more careful study is required. One may expand this list if necessary in applying this strategy to particular problems.

The second strategy derives from asking the question 'if such an interaction occurred between a ligand and a receptor, how many translational and rotational degrees of freedom would be frozen out in the process?'. In the CNS receptors, for example, only two features are needed to uniquely identify how a ligand would orient in the binding pocket (up to a mirror reflection): the distance from the phenyl ring to the basic amine, the angle of the lone pair on the amine to the plane of the ring, and the torsion angle of that lone pair to the plane of the ring.

The third strategy is justified by the reasons already described by Mayer et al.: if the active compounds are all binding in a common way to the same receptor, there must be some common binding geometry into which each compound can flex, i.e. the functional groups on the receptor responsible for binding possess (approximately) a unique spatial relationship. Of course, many physical processes can conspire to defeat such a strategy: there may be multiple binding modes, the functional group on the receptor may be highly flexible (e.g. a lysine), or the hydrogen bonds from the ligand to the receptor may in some cases, but not in other cases, form via an intermediate water. For these reasons, we do not demand that all actives fall into a common binding geometry; as long as 80% or more fall into a common pattern, one should continue the construction procedure.

One hitherto unexploited aspect of the method of Mayer et al. is that, in principle, it should allow one to identify multiple overlap regions, i.e. one may see some compounds cluster in one region of space and others in another. This would identify that one of these physical processes such as multiple binding modes is taking place. Figure 3 schematically shows how this may be detected.

The fourth strategy, the *selectivity principle*, may be viewed in two ways: one based on our knowledge of biochemistry and the other mathematical. Viewing this biochemically, we expect our 3D database queries to reflect properties of receptors, and we know that receptors must be exquisitely selective. Suppose, for example, our active analogs were all a series of hormones, and we found two patterns common to all of them, one which was rarely shared by molecules at random and another which was quite common to molecules found in the bloodstream. It makes intuitive sense to prefer the rarer, more selective pattern as our preferred 3D database query to find new

molecules to mimic the action of those hormones; it is unlikely that a hormone receptor would possess the second pattern, implying that the receptor would respond to a large proportion of the molecules to which it is exposed. We can sharpen this biochemical insight with the mathematical arguments of Karlin and Brendel [17]. They make the point that when structural patterns are found in a set of entities with common biological properties, one must be cautious about asserting the possibility of a cause-effect relationship between those structural patterns and those biological properties. In particular, one must assess how unusual those patterns are *relative to a population of like molecules*. Applying their ideas to our problem, take the case where all the active analogs are dipeptides. One structural pattern which we may find common to all these actives is that a positive-charge feature is 10–15 Å removed from a negative-charge feature. Among all molecules, this is a fairly rare and unusual pattern, *yet among dipeptides it is trivial*, and will be present in every dipep-

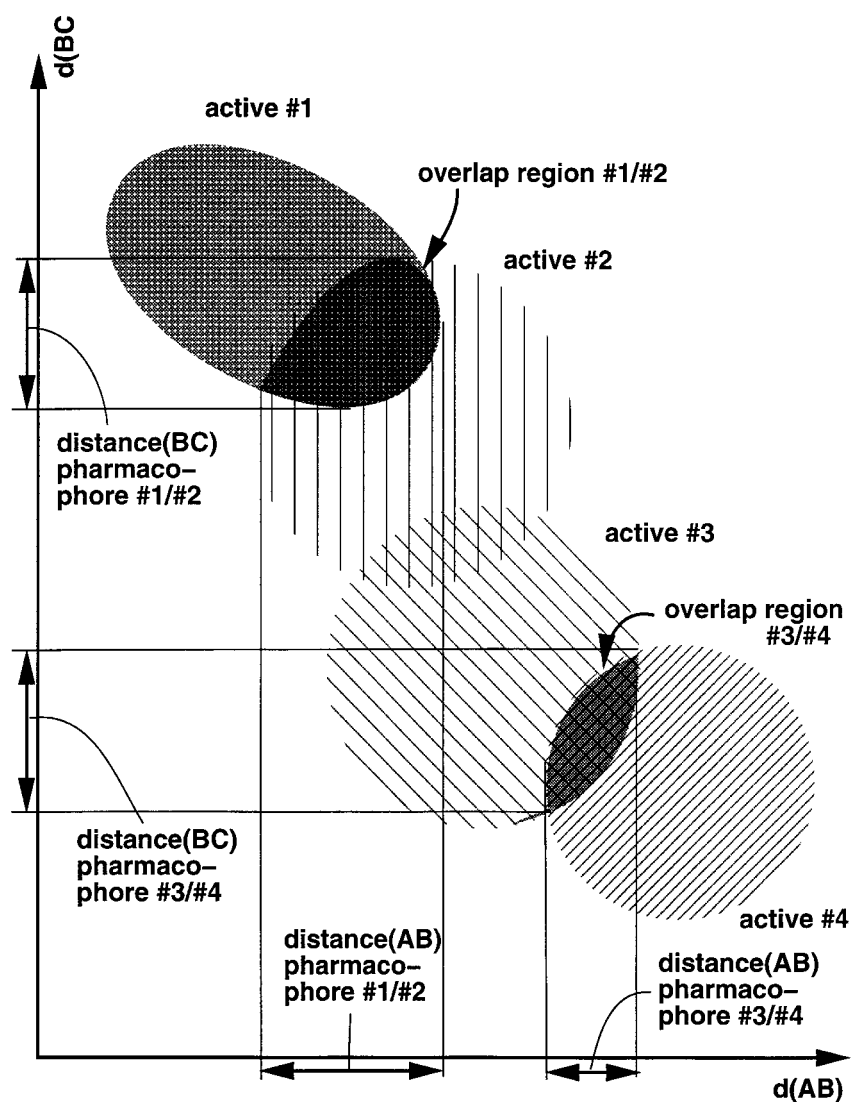
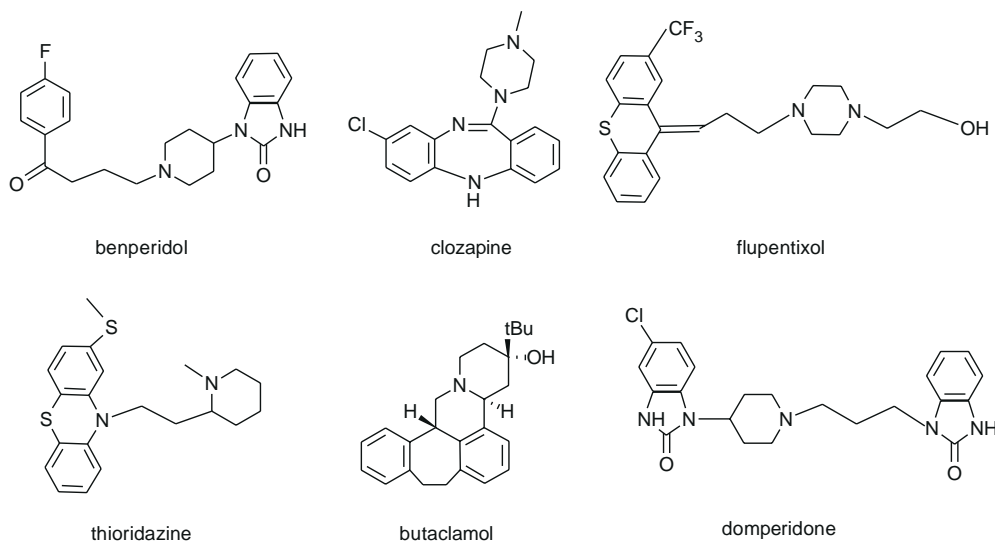
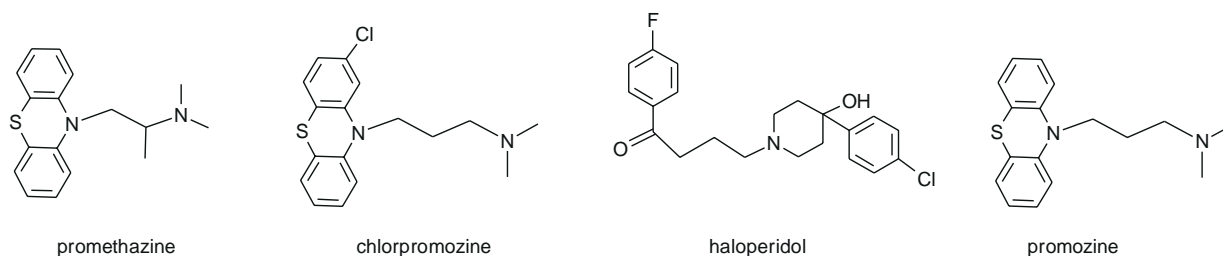


Fig. 3. Depiction of how the method of Mayer et al. may be used to detect the existence of receptor subtypes or multiple binding modes.

a



b

Fig. 4. D<sub>2</sub>-antagonist data set.

tide (the N- and C-termini). When judging selectivity, one must assess the selectivity of a pattern common to a set of actives against a population of like molecules. Furthermore, one can assess the statistical significance of a correlation between a common structural pattern and biological activity. If  $q$  is the likelihood that any molecule like the actives possesses that pattern, and  $N$  is the number of actives possessing that pattern, then the likelihood of that pattern appearing *by chance* among those actives is  $q^N$ . We will call this quantity our index of selectivity, and denote it by the variable  $S$ . Practically, we can compute the index of selectivity of a pattern of binding common to  $N$  active analogs by expressing it as a 3D database query and searching a database of drug-like molecules or a database of peptides. If  $q$  is the proportion of molecules from that database search that are returned as hits, then  $S = q^N$ . Thus, if we have 12 active analogs which are drug-like molecules sharing a structural pattern common to only 10% of the molecules in a database of drug-like molecules, the selectivity index is  $0.10^{12} = 10^{-12}$ . For convenience, we will often refer to  $-\log_{10}(S) = pS$ ; in this case,  $pS = 12$ . Such a pattern would be highly unlikely to correlate with biological activity by chance. This fourth strategy of preferring highly selective queries automatically minimizes the multiple-mapping problem. If the query represents a pattern that is highly unlikely to appear in any randomly chosen molecule, then it is especially un-

likely that that pattern will appear *twice* in the same molecule.

Values of  $q$ , the proportion of a database a given query hits, may be computed by performing a 3D database search, or by estimating the number of hits from a series of tabulated values, using a recently introduced inequality [18]

$$|AB(ablo, abhi)| \leq |AB(0, abhi)| + |AB(ablo, \infty)| - |AB(0, \infty)|$$

where  $AB(xylo, xyhi)$  refers to a distance constraint on two features A and B, constrained to lie from  $xylo \text{ \AA}$  to  $xyhi \text{ \AA}$ , and  $||$  denotes the number of hits returned by a query containing all the constraints within those brackets. When conformational flexibility is treated well, the inequality approaches an equality, allowing reasonably accurate estimates of  $q$ . That inequality as published was written only for dyad queries, but the mathematical arguments in that paper allow it to be generalized to a query with an arbitrary number of features.

The fifth strategy has purely practical motivations, to prune the number of possibilities that would need to be explored by applying all the earlier strategies simultaneously with four or five features (if  $F$  is the total number of features in our standard library and  $C$  is the total number of features we want in our final query, then the total number of possibilities is  $F^C$ ). Empirically, this strategy appears to work, and it has rarely been observed that

TABLE 1  
D<sub>2</sub>-ANTAGONIST DYADS

Dyad	Most selective distance range (Å)	q	N	pS
A-A	3.0–3.7	0.47	9	2.9
A-H	6.1–6.4	0.61	10	2.2
A-R	5.7–6.2	0.32	10	4.9
A-P	3.0–3.7	0.30	9	4.7
H-P	5.7–5.9	0.21	10	6.7
R-P	5.1–6.2	0.15	10	8.3

a high-ranking dyad does not appear in a high-ranking triad, when this procedure was begun in one case with two features and in another with three features. It is important to keep the number of possibilities in check, to ensure that we can expand our library of standard features when the need arises.

### Retrospective analyses for receptors of current interest

#### D<sub>2</sub>-antagonists

Begin with the 10 D<sub>2</sub>-antagonists shown in Fig. 4, taken from Seeman et al. [14]. Identify each of the six library features by one-letter codes: A – hydrogen-bond acceptor; D – hydrogen-bond donor; R – aromatic ring/pi-stacking; H – hydrophobe; P – group with a positive charge or capable of acquiring one at physiological pH; N – group with a negative charge or capable of acquiring one at physiological pH.

Only the dyads shown in Table 1 may be constructed, which hit all 10 D<sub>2</sub>-antagonists (here we consider only the simplest case of one distance constraint in the dyad). The value q was determined against the BioByte database. This database (formerly known as the ‘Pomona database’) is a collection of 26 000 drug-like compounds with some data on their biological activity, compiled by C. Hansch and

TABLE 2  
D<sub>2</sub>-ANTAGONIST TRIADS

Triad	Most selective distance ranges (Å)	q	N	pS
HP-A	H-P 5.5–6.2 P-A 3.0–5.4 H-A 3.5–8.9	0.16	9	7.2
HP-H	H <sub>1</sub> -P 5.5–6.2 H <sub>2</sub> -P 4.5–8.5 H <sub>1</sub> -H <sub>2</sub> 4.6–6.7	0.12	9	8.3
RP-A	R-P 5.1–6.2 R-A 3.1–8.3 P-A 3.5–4.9	0.10	9	8.9
RP-H	R-P 5.1–6.2 R-H 4.4–6.6 P-H 4.5–8.6	0.09	10	10.6
RP-R	R <sub>1</sub> -P 5.1–6.2 R <sub>2</sub> -P 3.8–7.9 R <sub>1</sub> -R <sub>2</sub> 4.6–6.0	0.004	10	13.7

A. Leo during the course of decades of QSAR studies. It is available from BioByte Corp., Claremont, CA, U.S.A.

The two dyad queries H-P and R-P stand out as being especially selective, with the likelihood that such patterns might appear by chance among 10 randomly chosen drug-like molecules being  $1.6 \times 10^{-7}$  and  $5.7 \times 10^{-9}$ , respectively. Hereafter, these dyads will be referred to as HP and RP (i.e. dyads with those two features and the distance constraints listed above). Checking how this dyad maps to each of the actives via graphical inspection, one sees that it maps uniquely on each, and consistently across the entire set (i.e. in accordance with our intuition). Either of these dyads represents the central features of most literature D<sub>2</sub>-antagonist pharmacophores [13], as well as those for almost any CNS receptor [19]. Advancing to triads, we will keep HP and RP as the most selective dyads, and construct 12 triad possibilities based on these dyads: HP-A, HP-D, HP-H, HP-P, HP-N, HP-R, RP-A, RP-D, RP-H, RP-P, RP-N, RP-R. We find that only five triads hit all 10 actives, HP-A, HP-H, RP-A, RP-H, RP-R. Applying the procedure of Mayer et al. to determine constraint values which will overlap the conformational space of each active, and choosing among any multiple solutions by selecting that combination of constraints which is the most selective, we obtain the triads shown in Table 2.

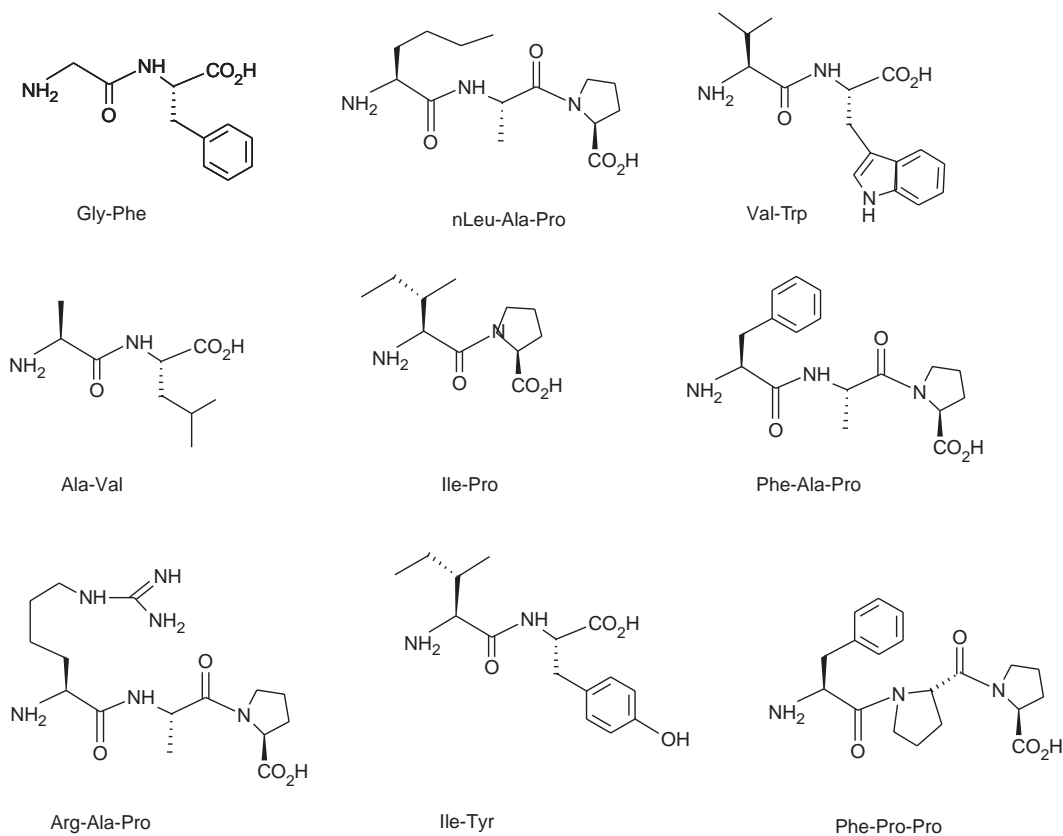
The three triads RP-A, RP-H and RP-R stand out as the most selective triads. Only RP-A and RP-R map the actives uniquely and consistently. We will refer to these triads as RPA and RPR in the following. Both are subsets of the previously cited literature D<sub>2</sub>-antagonist pharmacophores. These triads are highly selective queries; it probably suffices to terminate the progressive construction procedure at triads. It is instructive nonetheless to proceed one more step, to tetrads. Only the following tetrads will hit most of the actives: RPA-H, RPA-R, RPR-A, RPR-H. The procedure of Mayer et al. provides us with the tetrads shown in Table 3.

We can see that this procedure has produced two queries, each with two R features, one P, and one A. They

TABLE 3  
D<sub>2</sub>-ANTAGONIST TETRADS

Tetrad	Most selective distance ranges (Å)	q	N	pS
RPR-A	R <sub>1</sub> -R <sub>2</sub> 4.6–6.0 R <sub>1</sub> -P 5.1–6.2 R <sub>1</sub> -A 3.1–8.3 R <sub>2</sub> -P 3.8–6.2 R <sub>2</sub> -A 3.1–8.7 P-A 3.1–5.9	0.03	9	14.1
RPA-R	R <sub>1</sub> -R <sub>2</sub> 4.6–8.7 R <sub>1</sub> -P 5.1–6.2 R <sub>1</sub> -A 3.1–8.3 R <sub>2</sub> -P 3.7–7.0 R <sub>2</sub> -A 3.1–10.2 P-A 3.5–4.8	0.03	9	13.3

a



b

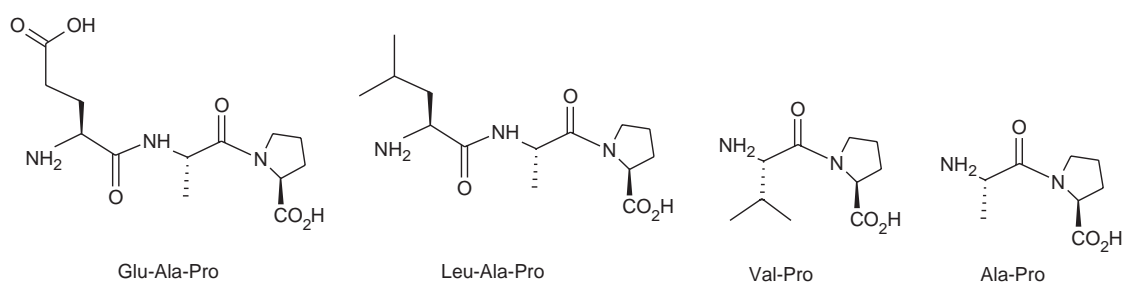


Fig. 5. ACE data set.

are not identical, but quite similar. If one composes a query combining the tightest constraints of each of these

pairs, one ends up with an RRPA query which hits 9 of the 10 actives, and has  $q=0.0018$ ,  $pS=15.7$ . However, the

TABLE 4  
ACE INHIBITOR DYADS

Dyad	Most selective distance range (Å)	N	q(BioByte)	pS(BioByte)	q(peptide)	pS(peptide)
H-N	4.1–4.7	11	0.06	13.4	0.71	1.8
D-P	5.8–6.1	13	0.06	15.9	1	0
N-P	5.2–5.6	11	0.007	23.7	0.93	0.3
A-H	4.2–4.6	13	0.77	1.5	0.83	1.0
H-P	4.4–5.0	11	0.19	7.9	0.70	1.7
A-D	3.4–3.5	13	0.17	10.0	1	0
D-D	5.7–6.0	13	0.10	13	1	0
D-H	4.0–4.4	12	0.36	5.3	0.83	1.0
D-N	5.3–5.6	13	0.02	22.1	1	0
A-P	5.9–6.1	13	0.11	12.5	1	0
A-N	4.4–4.7	13	0.04	18.2	1	0
A-A	3.3–3.4	13	0.27	7.4	1	0

TABLE 5  
ACE INHIBITOR TRIADS

Triad	Most selective distance ranges (Å)	N	q(peptide)	pS(peptide)
HN-P	H-N 4.1–4.7	11	0.24	6.8
	H-P 3.0–4.6			
	N-P 3.7–6.6			
HN-D	H-N 4.1–4.7	8	0.46	2.7
	H-D 5.2–6.1			
	N-D 5.3–6.0			
HN-A	H-N 4.1–4.7	7	0.27	4.0
	H-A 5.2–6.1			
	N-A 5.3–5.9			
HP-N	H-P 4.4–5.0	10	0.31	5.1
	H-N 3.2–4.5			
	P-N 5.1–7.4			

one active it does not hit is spiperone, the most active of the group, so this query is suspect.

#### ACE inhibitors

We will begin with 13 submillimolar peptides which were known to Ondetti and Cushman prior to their discovery of captopril, shown in Fig. 5. Our aim is to apply the strategies for constructing 3D database queries, to arrive at a query which ‘points in the right direction’, i.e. which retrieves known ACE inhibitors from the BioByte database.

One very important difference in this case will be that we evaluate the selectivity relative to a database of di- and tripeptides and *not* relative to the BioByte database. We will compile both  $q(\text{peptide})$  and  $q(\text{BioByte})$  to make clear the importance of this fact, and  $pS(\text{peptide}) = -\log(q(\text{peptide})^N)$  and  $pS(\text{BioByte}) = -\log(q(\text{BioByte})^N)$ . The most selective dyads shown in Table 4 hit most or all of the 13 actives.

Based on the selectivity against the peptide database, the HN dyad is the most selective; it hits the proline ring and the carboxy terminus in most of the actives. This is

TABLE 6  
 $\beta_2$ -ANTAGONIST DYADS

Dyad	Most selective distance range (Å)	q	N	pS
A-H	3.8–3.9	0.40	5	2.0
D-R	6.0–6.1	0.08	5	5.4
H-D	6.1–6.2	0.28	5	2.8
H-P	7.0–7.1	0.16	5	4.1
P-R	5.1–5.2	0.09	5	5.3

an example of a rare case when the 3D query picks up a topological feature common to almost all the actives: a C-terminal proline. This feature is common to most known ACE inhibitors as well, so this dyad indeed is ‘pointing in the right direction’. Notice that had we used the BioByte database for measuring the selectivity, the HN dyad would *not* have been near the top; the NP dyad would have been identified as the most selective. Yet having that pattern appear in a set of di- and tripeptides is almost a trivial occurrence, since that hits the C- and N-termini. Since the HN and HP dyads stand far above all the others, we will proceed to build triads only with these. Only the triads depicted in Table 5, constructed from those dyads, hit most of the actives.

Each of these queries retrieves at least one known ACE inhibitor from the BioByte database. Two percent of the hits of HN-P are ACE inhibitors (1, delapril out of 49 hits); HN-D, 2% (captopril and delapril, out of 97 hits); HN-A, 1% (captopril, alacepril, delapril out of 236 hits), and they all express a similar pattern: the proline ring forms the basis, and the next feature is the amide bond or N-terminus if a dipeptide. This amide bond or N-terminus is picked up either as a donor, an acceptor, or as an N-terminal positively charged group. This is consistent with the generally accepted pattern of binding for ACE inhibitors (except that what we pick up as a donor is in fact binding to a Zn).

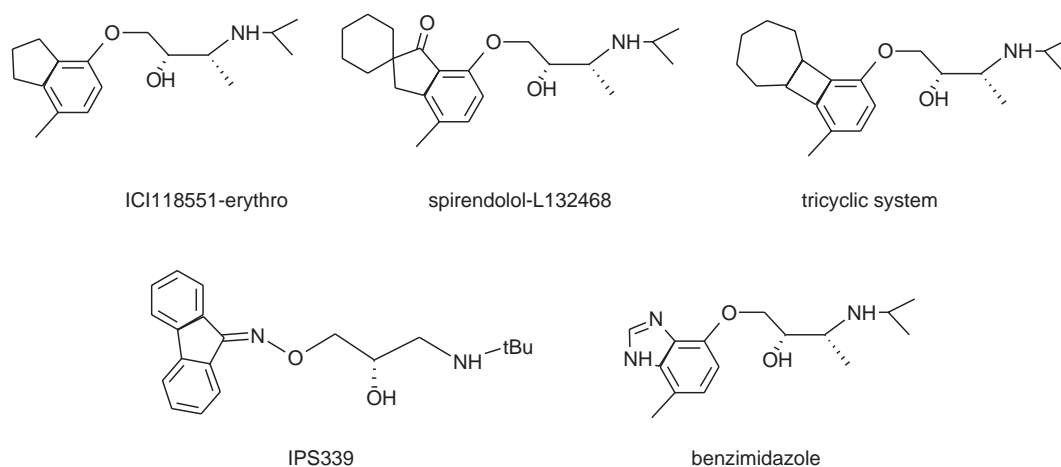


Fig. 6.  $\beta_2$ -antagonist data set.

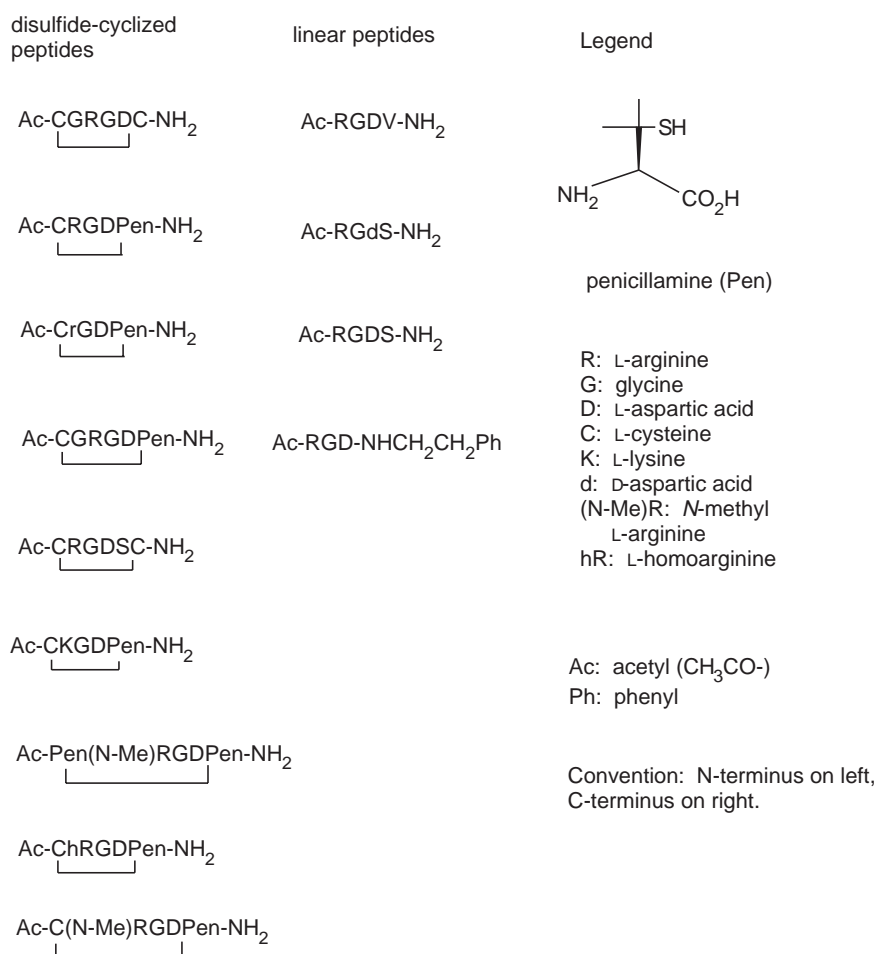


Fig. 7. Fibrinogen antagonist ('RGD') data set.

TABLE 7  
β<sub>2</sub>-ANTAGONIST TRIADS

Triad	Most selective distance ranges (Å)	N	q(BioByte)	pS(BioByte)
PR-A	P-R 5.1–5.2	5	0.02	8.2
	P-A 3.0–4.1			
	R-A 3.0–4.1			
PR-D	P-R 5.1–5.2	5	0.015	9.1
	P-D 3.0–3.8			
	R-D 3.0–3.8			
PR-H	P-R 5.1–5.2	4	0.01	8.1
	P-H 6.9–7.5			
	R-H 6.9–7.5			
HP-A	H-P 7.0–7.1	5	0.02	8.6
	H-A 6.6–6.9			
	P-A 3.6–3.8			
HP-D	H-P 7.0–7.1	5	0.025	8.0
	H-D 6.6–6.9			
	P-D 3.6–3.8			
HP-H	H <sub>1</sub> -P 7.0–7.1	4	0.04	5.6
	H <sub>1</sub> -H <sub>2</sub> 3.2–3.3			
	P-H <sub>2</sub> 6.7–6.9			
HP-R	H-P 7.0–7.1	4	0.025	6.4
	H-R 3.1–3.4			
	P-R 6.4–7.1			

*β<sub>2</sub>-antagonists*

The pattern of binding of β<sub>2</sub>-antagonists has been well characterized by the molecular biological studies of Strader et al. [20]. They have demonstrated that the phenyl ring interacts with Phe<sup>290</sup>; the catechol binds to Ser<sup>204</sup> and Ser<sup>207</sup>; and the basic nitrogen binds to Asp<sup>113</sup>. Taking five β<sub>2</sub>-antagonists from Hansch et al. [21], as shown in Fig. 6, the most selective dyads that emerged are shown in Table 6.

Of these dyads, only the HP and PR dyads map con-

TABLE 8  
RGD ANTAGONIST DYADS

Dyad	Most selective distance range (Å)	N	q(peptide)	pS(peptide)
P-N	11.9–12.2	12	0.13	10.6
A-P	10.3–10.4	15	0.58	3.5
H-N	9.3–9.4	12	0.58	2.8
H-P	3.7–3.9	11	0.61	2.4
A-H	9.2–9.3	16	0.81	1.5
D-N	4.1–4.3	14	0.78	1.5
D-D	10.4–10.5	16	0.81	1.5
D-H	4.5–4.6	15	0.83	1.2

TABLE 9  
RGD ANTAGONIST TRIADS

Triad	Most selective distance ranges (Å)	N	q(peptide)	pS(peptide)
PN-A	P-N 11.9–12.2	9	0.08	9.9
	P-A 7.8–9.6			
	N-A 4.7–6.6			
PN-D	P-N 11.9–12.2	9	0.10	8.9
	P-D 5.9–7.6			
	N-D 5.6–7.2			

sistently (to the phenyl ring and the basic nitrogen), so we construct triads from them (see Table 7). The most selective triad is PR-D. This result is consistent with the conclusions based on molecular biological data. It is interesting to note that PR-D returns 363 hits, 88 of which are  $\beta_2$ -antagonists (24%). PR-A returns 571 hits from the BioByte database, 19 of which are  $\beta_2$ -antagonists (3%).

#### Fibrinogen antagonists ('RGD' binding site)

These hold promise for thrombolytic therapy, as one of the crucial steps in platelet aggregation is the binding of fibrinogen to the membrane-bound protein GP IIb/IIIa. The key part of fibrinogen involved in this interaction is an Arg-Gly-Asp (RGD) sequence on one of the loops. A series of cyclic peptides containing this RGD sequence have been synthesized and evaluated for their biological activity [22]. Sixteen cyclic and acyclic fibrinogen antagonists are shown in Fig. 7.

The dyads in Table 8 emerged as the most selective. Note that the PN dyad stands out far above the rest; this is a more selective version of the dyad query evidently used by Merck in a 3D search of their sample collection (they looked for basic amines separated from a carboxylate by 10–20 Å) [23].

The triads one forms from PN are shown in Table 9. Both of these triads pick up as an additional feature an element of an amide bond, in one case the oxygen and in the other the NH.

#### Testing of assumptions

G. Smith (personal communication) has suggested that any such pharmacophore identification procedure should have a smooth 'dose-response', i.e. that slight variations

TABLE 10  
EFFECTS OF CONFORMATIONAL PERTURBATIONS ON D<sub>2</sub>-ANTAGONIST DYADS

Dyad	Most selective distance range (Å)		
	Original	Perturb 0.1 Å	Perturb 1.0 Å
A-A	3.0–3.7	3.0–3.7	3.9–4.5
A-H	6.8–7.0	6.8–6.9	4.7–4.8
A-P	3.0–3.7	3.0–3.7	3.9–4.5
A-R	5.7–6.2	5.7–6.2	5.6–5.8
H-P	4.9–5.0	4.9–5.0	4.9–5.0
P-R	5.1–6.2	5.2–6.2	5.6–5.8

in the input yield at most slight changes in the output. Furthermore, if this method claims to develop 3D database queries which are physically relevant, it is important that changing various arbitrary assumptions do not fundamentally alter the results. We will consider specific aspects here.

#### Variations in conformational analysis

Inevitably, the quality of the results of this approach to developing 3D database queries will depend on the quality of the conformational analysis used in exploring the conformational space of each molecule. A simple computational experiment is to take the original set of conformers used in the D<sub>2</sub>-antagonist example, and perform three transformations: (1) rotate each molecule by 30° about the z-axis; (2) randomly permute the coordinates of each atom by an amount on each axis that is equally likely in the range [+0.1 Å, -0.1 Å]; and (3) randomly permute the coordinates by an amount [+1.0 Å, -1.0 Å]. For case (1), we expect absolutely no change up to the precision of the computer. For case (2), we expect at most that some of the distance ranges may increase or decrease by 0.1 Å, which may produce slight changes in the selectivity rankings. For case (3), we should anticipate that we no longer will get meaningful results.

For case (1), we see no change, as expected. Table 10 shows the dyads that emerged from the original analysis, along with those from case (2), 'perturb 0.1 Å', and case (3), 'perturb 1.0 Å'. With the 0.1 Å perturbation, we would have still been led to the same conclusions; note that all the differences in dyads are less than or equal to 0.1 Å. At the level of 1.0 Å, the results begin to show sensitivity to the conformational analysis.

TABLE 11  
EFFECTS OF DATABASE CHOICE IN D<sub>2</sub>-ANTAGONIST DYAD RANKINGS

Dyad	q(BioByte)	pS(BioByte)	rank(BioByte)	q(NCI)	pS(NCI)	rank(NCI)
R-P	0.15	8.8	1	0.09	10.5	1
H-P	0.21	6.7	2	0.14	8.5	2
A-R	0.32	4.9	3	0.34	4.7	4
A-P	0.30	4.9	4	0.15	7.4	3
A-A	0.47	2.9	5	0.46	3.0	5
A-H	0.61	2.2	6	0.59	2.4	6

TABLE 12  
EFFECTS OF DATA SET SELECTION ON D<sub>2</sub>-ANTAGONIST DYADS

Dyad	Original		Set A		Set B		Set C	
	Most selective distance range (Å)	pS	Distance range (Å)	pS	Distance range (Å)	pS	Distance range (Å)	pS
A-A	3.0–3.7	2.9	3.5–3.7	0.7	3.0–3.7	0.9	3.0–3.6	0.8
A-H	6.1–6.4	2.2	6.9–7.0	0.5	6.8–6.9	0.5	6.8–7.0	0.5
A-P	3.0–3.7	4.7	3.5–3.7	3.6	3.0–3.7	4.2	3.0–3.6	0.2
A-R	5.7–6.2	4.9	5.8–5.9	1.2	6.2–8.0	1.4	6.1–6.3	1.3
H-P	5.7–5.9	6.7	5.6–5.8	4.0	5.0–5.1	4.0	5.4–5.5	4.0
R-P	5.1–6.2	8.3	5.4–5.9	4.7	5.4–5.5	4.8	6.1–6.2	5.0

Sets A, B, and C are randomly selected subsets of the original data set.

#### Use of other drug-like databases other than BioByte

Our selectivity measure depends critically on  $q(\text{BioByte})$ , the proportion of hits a query returns from the BioByte database. If our conclusions are to represent something physically meaningful, they should not depend on which database of drug-like molecules one selects (as long as the database represents a wide, balanced spectrum of drug-like molecules, which is *not true* for most corporate drug databases). Repeating our analysis of D<sub>2</sub>-antagonists with the NCI database (this is a collection of drug-like molecules that have been tested for carcinogenicity, and which have CAS numbers; it is available from the National Cancer Institute, Frederick, MD, U.S.A.), we obtain the dyad rankings shown in Table 11. We see that the selectivity value varies, but the rankings of the most selective dyads do not.

#### Influence of data set selection

Similarly, if our resulting database queries are to represent something physical, they should not depend significantly on the somewhat arbitrary decisions we make in choosing molecules for our data sets. Creating three different data sets of six molecules, each randomly chosen from the original 12, we obtain the dyad rankings shown in Table 12.

This shows that the rankings of the most selective dyads are not significantly affected by the data set selection. The smaller data sets will tend to produce queries

TABLE 13  
EFFECTS OF NOISE ON D<sub>2</sub>-ANTAGONIST DYADS

Dyad	Original		New	
	Most selective distance range (Å)	N	Most selective distance range (Å)	N
R-P	5.1–6.2	10	5.1–6.2	10
H-P	5.7–5.9	10	4.7–5.2	12
A-R	5.7–6.2	10	5.7–6.2	10
A-P	3.0–3.7	9	3.0–3.7	11
A-A	3.0–3.7	9	3.0–3.7	11
A-H	6.1–6.4	10	5.4–6.0	12

Three molecules randomly chosen from the BioByte database have been added to the original data set.

that are *too* selective (i.e.  $q$  too small), and due to the smaller  $N$  this makes all the selectivity rankings less significant; it is important to find a reasonably sized population of actives to ensure our resulting query spans the space of all possibilities.

#### Influence of random molecules added to a set of actives

It is not infrequently the case that biological assays will pick up a molecule whose high potency is an artifact. Similarly, one often encounters cases where not all molecules are binding in a common fashion. As mentioned earlier, ideally any method for building 3D database queries would detect molecules which do not fit a pattern common to the rest of the actives, as would be true in both of these cases. We can construct test cases to test the ability of these strategies to do exactly that, by adding totally random molecules to some of the above data sets. Let us add three molecules randomly chosen from the BioByte database to the D<sub>2</sub>-antagonist data set described earlier. Repeating our analysis, we obtain the dyads shown in Table 13.

#### Totally random sets of molecules

Up to this point, we have used the selectivity measure  $pS$  merely to rank different possibilities. It would be useful to calibrate this quantitatively, to understand at what values of that measure can we conclude that we have

TABLE 14  
DYADS RESULTING FROM DATA SET 'a' OF 10 RANDOMLY CHOSEN MOLECULES

Dyad	Most selective distance range (Å)	N	$q(\text{BioByte})$	$pS(\text{BioByte})$
A-D	3.6–6.1	8	0.50	2.4
A-H	7.0–7.5	8	0.52	2.2
A-R	3.6–6.3	8	0.56	2.0
A-A	3.6–6.6	8	0.64	1.5
H-H	4.0–4.7	8	0.44	2.9
D-R	3.7–7.0	7	0.34	3.2
H-P	7.0–7.5	7	0.18	5.3
H-R	5.5–6.3	7	0.34	3.2
P-R	3.9–7.1	7	0.18	5.2
D-H	2.8–3.3	7	0.32	3.4

TABLE 15  
DYADS RESULTING FROM DATA SETS 'b' AND 'c', TWO  
DISTINCT SETS OF 10 RANDOMLY CHOSEN MOLECULES

Data set	Dyad	Most selective distance range (Å)	N	q(BioByte)	pS(BioByte)
b	A-D	2.9–6.1	5	0.52	1.4
c	A-D	3.5–5.6	8	0.50	2.4
c	A-A	4.6–5.2	8	0.48	2.5
c	H-A	3.0–4.3	6	0.81	0.5
c	H-D	3.0–4.9	6	0.56	1.5

achieved a meaningful result. One simple way to do this is to apply these strategies to a set of randomly chosen molecules. We have taken three different sets of randomly chosen molecules from the BioByte database, labeled a, b, and c, each set consisting of 10 molecules. Treating each set on its own, and searching for the most selective dyads as before, we obtain those depicted in Tables 14 and 15. Notice that the selectivity measure pS never gets significantly larger than 5 with these random data sets. This suggests that any database queries we construct with selectivities of that order are not that significant.

## Conclusions

We have outlined a set of strategies for constructing 3D database queries from a set of active molecules, whose conformational space we have fully explored. These strategies allow one to construct many possibilities, based on a standard library of six features common to most ligand–receptor interactions. The geometrical pattern of these features common among most of the actives is determined by the method of Mayer et al. Finally, these possibilities are ranked by their *selectivity*,  $pS = -\log(q^N)$ , where  $q$  is the proportion of molecules returned as hits from a database of molecules structurally similar to the actives and  $N$  is the number of actives which possess this pattern. Applications of this method were presented, in cases both where the active leads were predominantly peptidic and where they were more typically 'drug-like'. We have shown how the 3D database query either represents a pattern of binding similar to that determined by other methods, or could have led to the discovery of leads that were found by other methods. We have shown that this method is robust relative to variations in the input, data set selection, and the introduction of bad or misleading data. We have also seen that this method provides quantitative feedback when the input data are totally meaningless, in that very small values of pS result.

## Acknowledgements

This analysis was performed with custom-written programs for the procedure of Mayer et al., in conjunction

with the conformational analysis and flexible database searching tools contained in Catalyst v. 2.1. [24]. The author thanks P.W. Sprague for suggesting the ACE inhibitor case, and making available the early data of Ondetti and Cushman.

## References

- a. Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R. and Ferrin, T.E., *J. Mol. Biol.*, 161 (1982) 269.
- b. DesJarlais, R.L., Sheridan, R.P., Dixon, J.S., Kuntz, I.D. and Venkataraghavan, R., *J. Med. Chem.*, 29 (1986) 2149.
- Gund, P., *Annu. Rep. Med. Chem.*, 14 (1979) 299.
- Brint, A.T. and Willett, P., *J. Mol. Graph.*, 5 (1987) 49.
- VanDrie, J.H., Weininger, D. and Martin, Y.C., *J. Comput.-Aided Mol. Design*, 3 (1989) 255.
- Sheridan, R.P., Nilakantan, R., Rusinko, A., Bauman, N., Haraki, K.S. and Venkataraghavan, R., *J. Chem. Inf. Comput. Sci.*, 29 (1989) 255.
- a. Christie, B.D., Henry, D.R., Güner, O.F. and Mook, T.E., *Online Inf.*, 90 (1990) 137.
- b. Mook, T.E., Henry, D.R., Ozkabak, A.G. and Alamgir, M., *J. Chem. Inf. Comput. Sci.*, 34 (1994) 184.
- Murrall, N.W. and Davies, E.K., *J. Chem. Inf. Comput. Sci.*, 30 (1990) 312.
- Van Drie, J.H., Berezin, S. and Ku, S.-L., Abstracts for the ACS National Meeting, Spring, 1992, CINF 023. No detailed description of the Catalyst 3D database searching software has been published.
- Hurst, T., *J. Chem. Inf. Comput. Sci.*, 34 (1994) 190. No further details have been published on the Unity 3D database searching software.
- Marshall, G.R. and Cramer, R.D., *Trends Pharmacol. Sci.*, 9 (1988) 285.
- Good, A.C. and Mason, J.S., In Lipkowitz, K.B. and Boyd, D.B. (Eds.) *Reviews in Computational Chemistry*, Vol. 7, VCH, New York, NY, U.S.A., 1996, pp. 67–117.
- Mayer, D., Naylor, C.B., Motoc, I. and Marshall, G.R., *J. Comput.-Aided Mol. Design*, 1 (1987) 3.
- Ondetti, M.A., Cushman, D.W. and Rubin, B., In Bindra, J.S. and Lednicer, D. (Eds.) *Chronicles of Drug Discovery*, Vol. 2, Wiley, New York, NY, U.S.A., 1983, pp. 1–31.
- Seeman, P., Watanabe, M., Grigoriadis, D., Tedesco, J.L., George, S.R., Svensson, U., Nilsson, J.L. and Neumeyer, J.L., *Mol. Pharmacol.*, 28 (1985) 391.
- a. Bartlett, P.A., Shea, G.T., Telfer, S.J. and Waterman, S., In Roberts, S.M. (Ed.) *Molecular Recognition: Chemical and Biological Problems*, Vol. 78, Royal Society of Chemistry, London, U.K., 1989, pp. 182–192.
- b. Lauri, G. and Bartlett, P.A., *J. Comput.-Aided Mol. Design*, 8 (1994) 51.
- Van Drie, J.H. and Martin, Y.C., Workshop on 3D databases at the Crystallography and Drug Design Conference, Erice, Italy, June 1989.
- Karlin, S. and Brendel, V., *Science*, 257 (1992) 39.
- Van Drie, J.H., *J. Comput.-Aided Mol. Design*, 10 (1996) 623.
- Lloyd, E.J. and Andrews, P.J., *J. Med. Chem.*, 29 (1986) 453. Note that, while we identify the basic amine as a 'positive-charge' group, they identify it as an H-bond donor. It is difficult to say precisely what the protonation state of that nitrogen is when bound to the receptor.

- 20 Strader, C.D., Sigal, I.S. and Dixon, R.A.F., *FASEB J.*, 3 (1989) 1825.
- 21 Hansch, C., Sammes, P.G. and Taylor, P.G. (Eds.) *Comprehensive Medicinal Chemistry*, Pergamon, Oxford, U.K., 1990.
- 22 Samanen, J., Ali, F., Romoff, T., Calvo, R., Sorenson, E., Vasko, J., Storer, B., Berry, D., Bennett, D., Strohsacker, M., Powers, D., Stadel, J. and Nichols, A., *J. Med. Chem.*, 34 (1991) 3114.
- 23 Hartman, G.D., Egbertson, M.S., Halczenko, W., Laswell, W.L., Duggan, M.E., Smith, R.L., Naylor, A.M., Manno, P.D., Lynch, R.J., Zhang, G., Chang, C.T.-C. and Gould, R.J., *J. Med. Chem.*, 35 (1992) 4640.
- 24 The Catalyst software is available from Molecular Simulations Inc., San Diego, CA, U.S.A. Conformers were generated using the so-called 'fast' methodology (torsion-angle driving with energy-

minimization post-processing) with an 8.0 kcal/mol cutoff. Up to 150 conformers/molecule were generated. The flexible database searching used does not perform any on-the-fly manipulation of the conformers, but rather relies strictly on the conformers stored in the database. There is some difference between the way in which Catalyst defines features and the way in which the custom software implementing the procedure of Mayer et al. defines them, especially for the 'hydrophobic' feature. However, in principle, one should be able to reproduce this analysis with any method of conformational analysis which effectively explores all energetically reasonable regions of conformational space, and any database search system which allows one both to express the concepts we discuss here, and which takes into account conformational flexibility in analyzing a molecule for possible hits.