Proc. Natl. Acad. Sci. USA
Vol. 86, pp. 8165–8169, October 1989
Pharmacology

Searching for pharmacophores in large coordinate data bases and its use in drug design

(topographic/three-dimensional substructure/CONCORD)

ROBERT P. SHERIDAN, ANDREW RUSINKO III, RAMASWAMY NILAKANTAN, AND R. VENKATARAGHAVAN

Medical Research Division, Lederle Laboratories, American Cyanamid, Pearl River, NY 10965

Communicated by F. W. McLafferty, May 25, 1989 (received for review February 25, 1989)

ABSTRACT **Pharmacophores, three-dimensional arrangements of chemical groups essential for biological activity, are being proposed in increasing numbers. We have developed a system to search data bases of three-dimensional coordinates for compounds that contain a particular pharmacophore. The coordinates can be derived from experiment (e.g., Cambridge Crystal Database) or be generated from data bases of connection tables (e.g., Cyanamid Laboratories proprietary compounds) via the program CONCORD. We discuss the results of searches for three sample pharmacophores. Two have been proposed by others based on the conformational analysis of active compounds, and one is inferred from the crystal structure of a protein–ligand complex. These examples show that such searches can identify classes of compounds that are structurally different from the compounds from which the pharmacophore was derived but are known to have the appropriate biological activity. Occasionally, the searches find bond "frameworks" in which the important groups are rigidly held in the proper geometry. These may suggest new structural classes for synthesis.**

The function of rational drug design is to relate chemical structure to a specific biological activity and then to use the relationship to select existing compounds for testing or to suggest new compounds for synthesis. Now that three-dimensional molecular modeling is widely used, "pharmacophore" models for a variety of biological activities are appearing in the literature in increasing numbers. The concept of a pharmacophore, although based on many simplifying assumptions, is a useful one: a particular spatial arrangement of chemical groups (usually atoms), common to all active molecules, that is recognized by a single receptor. Pharmacophores can be inferred from the structure of ligand–receptor complexes or, when the structure of the receptor is not known, can be deduced by conformational analysis of active compounds, as in the "active analog" approach (1–3). A pharmacophore is usually expressed in three-dimensional terms (distances, angles, volumes), and its atoms may be described by a physical property (cation, hydrogen bond donor, etc.). Often, a pharmacophore will contain one or more "dummy" atoms, which are used to define a geometric entity (centroid of a ring, a lone pair direction, etc.).

More than 10 years ago, Gund (4, 5) pointed out how, once a pharmacophore is defined for a particular biological activity, one could gain further insight by finding structures that contain the pharmacophore but are structurally different from the compounds from which the pharmacophore was derived. He described one of the first methods (MOLPAT) to search for pharmacophores in data bases of three-dimensional coordinates. At that time, the usefulness of such searches was limited since the three-dimensional data bases available (crystal coordinates or computer-generated conformations of a few compounds) were small. Now, however, the data base of crystal coordinates maintained by Cambridge (6) has >60,000 entries. Moreover, practical methods to transform large data bases of connection tables to coordinate form (7) have recently become available. Also, computer algorithms for doing three-dimensional substructure searches (8–10) have improved.

Searching for pharmacophores in large data bases has now become a practical way to exploit geometric structure–activity information. In this paper, we discuss the results of searching for three sample pharmacophores. Our primary goal is to demonstrate the type of insight that can come from pharmacophore searches.

METHODS

We have been working with two large data bases of coordinates: CCD, a subset of the Cambridge Crystal Database (6), and CLFIL, a three-dimensional version of the CL File, the set of proprietary structures from American Cyanamid. The 29,828 structures in CCD are those for which we could relate the coordinate and connection table information provided by the Cambridge Crystallographic Data Centre. The 223,988 structures in CLFIL were generated from a MACCS data base by using CONCORD. [MACCS (Molecular ACCess System) is the tradename for a chemical data base management system supplied by Molecular Design Ltd., San Leandro, CA.] CONCORD is a rule-based system that generates a single set of three-dimensional coordinates from a connection table that contains only "organic" atoms. Structures in either data base are stored as a set of atom types and coordinates for nonhydrogen atoms. The atom type consists of five fields: element, number of neighbors (i.e., bonded atoms), the number of pi electrons, the number of attached hydrogens at physiological pH, and formal charge. Four types of dummy atoms were added: D5 and D6 (the centroids of planar 5- and 6-membered rings), DP (along the perpendiculars to these rings), and DL (attached to each heteroatom along the vector sum of the bonds from its neighbors). These are equivalent to the types of dummy atoms most often used in describing pharmacophore geometries. Details of how the data bases were generated will be given elsewhere (7).

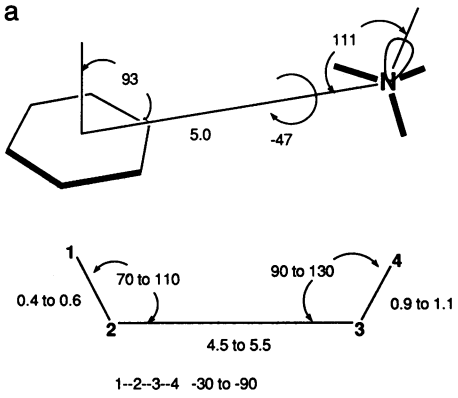
We have developed a system 3DSEARCH (8) that allows a user to define a three-dimensional substructure "query," a set of atoms and a description of the spatial relationship between them, and to search for it in large data bases of coordinates. The user can specify upper and lower bounds to selected distances, angles, and dihedral angles. Excluded volumes (regions that are not allowed to be occupied) can also be taken into account. We can define several possible atom types for each query atom, making it possible to specify "generic" types (cations, hydrogen-bond donors, etc).

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: CNS, central nervous system; ACE, angiotensin-converting enzyme.

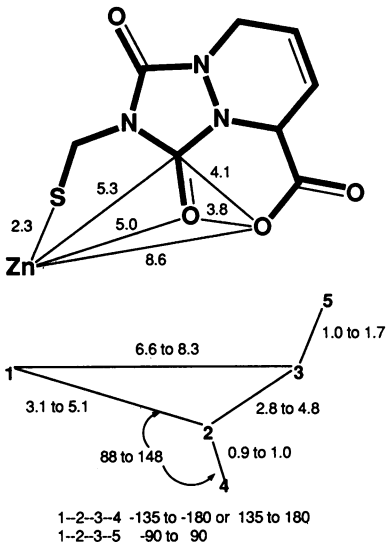
3DSEARCH uses a two-part search strategy: a rapid screen using a key-search algorithm and a slower atom-by-atom geometric search on the structures that pass the screen (typically a few percent of the original list). The keys are descriptors defined by pairs of atoms and the distances between them. We applied an "inverted" key-indexing scheme, in contrast to the "direct" scheme described by others (10) using similar keys. In an inverted key scheme, the

a



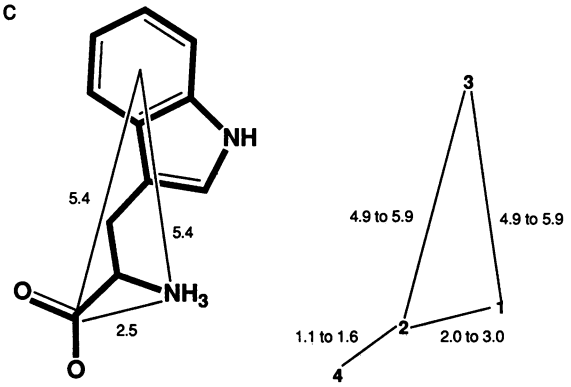
Atom	Type	Element	Neighbors	Pi	H's	Chg
1	1	DP	0	0	0	0
2	1	D6	0	0	0	0
3	1	N	3	0	1	1
4	1	DL	0	0	0	0

b



Atom	Type	Element	Neighbors	Pi	H's	Chg
1	1	S	1	0	Wi	Wi
	2	O	1	0	0	-1
	3	O	1	1	0	-1
2	1	N	2	1	0	0
	2	N	1	2	0	0
	3	O	2	0	0	0
	4	O	1	1	0	0
	5	F	1	0	0	0
3	1	S	1	0	0	-1
	2	O	1	0	0	-1
	3	O	1	1	0	-1
4	1	DL	0	0	0	0
5	1	C	3	1	0	0
	2	P	Wi	Wi	0	0
	3	S	Wi	Wi	0	0

c



Protein volume represented by 122 outriggers.

Atom	Type	Element	Neighbors	Pi	H's	Chg
1	1	N	1	0	3	1
	2	N	2	0	2	1
	3	N	3	0	1	1
	4	N	4	0	0	1
	5	N	3	1	0	1
	6	P	4	0	0	1
	7	S	3	0	0	1
2	1	C	3	1	0	0
	2	N	Wi	1	0	0
	3	S	Wi	Wi	0	0
	4	P	Wi	Wi	0	0
3	1	D5	0	0	0	0
	2	D6	0	0	0	0
4	1	O	1	0	0	-1
	2	O	1	1	0	-1

FIG. 1. Pharmacophores and queries derived from them. In the queries, all distances are given in angstroms and all angle/dihedral constraints are given in degrees. The distances to dummy atoms (D) are determined by the arbitrary distance we used to generate dummy atoms in the coordinate data bases (ref. 7). In each case the possible allowed atom types for each of the query atoms are shown. A "Wi" ("wild card") means any value is allowed. Pi, pi electrons; H's, attached hydrogens at physiological pH; Chg, formal charge; Neighbors, bonded atoms. For explanations of D5, D6, DP, and DL, see *Methods*. (*a*) CNS pharmacophore as proposed by Lloyd and Andrews (12). Shown on the top is the relationship of a basic tertiary amine to the phenyl ring in their "common model." Below is the query derived from this relationship. Atom 2 is the centroid of a flat 6-membered ring, atom 3 is the basic amine, atom 1 indicates the perpendicular to the ring, and atom 4 indicates the direction of the long pair on the amine. (*b*) The angiotensin-converting enzyme inhibitor pharmacophore proposed by Mayer *et al.* (14). A structure of a semirigid inhibitor (15), with distances indicating the relationship between essential atoms, including the enzyme-bound Zn, is shown at the top. Below this is shown the pharmacophore derived from 10 distance geometry solutions of the inhibitor given those distances. Atom 1 is a potential Zn ligand (sulfhydryl or carboxylate oxygen), atom 2 is a neutral H-bond acceptor, atom 3 is an anion (deprotonated sulfur or charged oxygen), atom 4 indicates the direction of a hydrogen bond to atom 2, and atom 5 is the central atom of a carboxylate, sulfate, or phosphate of which atom 3 is an oxygen or an unsaturated carbon when atom 3 is a deprotonated sulfur. (*c*) The tryptophan repressor pharmacophore. The relationship between the amino N, the carboxyl C, and the six-membered ring in the crystal structure of the tryptophan–tryptophan repressor complex (18) is shown on the left; on the right is the query. Atom 1 is a cation; atom 2 is the center of a carboxylate, phosphate, sulfate, etc.; atom 3 is the centroid of a planar 5- or 6-membered ring; and atom 4 is an anionic oxygen bonded to atom 2.

search time is proportional to the number of distances defined in the query and proportional to the number of distinct keys in the data base. The latter is roughly proportional to the logarithm of the number of structures in the data base. A typical key search over a data base of >200,000 structures takes about 30 CPU (central processing unit) sec on our VAX 8650. We modified the geometric search algorithm of Ullman (11) to take constraints other than distances (i.e., angle/dihedral angles and excluded volumes) into account. The geometric search takes 0.04–0.10 sec per structure, depending on the number of constraints. Thus, the overall search time is usually <5 min for typical pharmacophores. Details of the key and geometric search algorithms and the results of timing studies will be reported elsewhere (8).

RESULTS

The three sample pharmacophores we have chosen are shown in Fig. 1. How upper and lower bounds in a query are set around the "ideal" pharmacophore is always case-dependent and somewhat arbitrary. As in any kind of data base search, one must compromise between making the query too narrow, in which case many "interesting" structures are missed, or too broad, in which case too many "uninteresting" structures are found. (In our case, "uninteresting" structures will not make a convincing superposition with the ideal pharmacophore.) We believe we have made reasonable choices in the queries discussed here.

When the query contains a reference to absolute chirality (e.g., dihedral angles or excluded volume), we report the results as the union of the results of two searches: one over the original coordinates in the data base and one over their mirror reflections. This is done for two reasons. First, the absolute chirality of structures in CLFIL is arbitrary, having been built from connection tables that for the most part lack stereochemical information. Also, the absolute chirality of crystal coordinates in CCD is not always correct. Second, for many applications a mirror reflection is equally valuable, since it can represent the alternative enantiomer of a chiral compound or an alternative conformation of a nonchiral compound.

Lloyd and Andrews (12) have demonstrated that a variety of drugs with central nervous system (CNS) effects can attain a common orientation of a nitrogen and its lone pair relative to an aromatic ring. Their "common model" and our translation of this model into a query are shown in Fig. 1*a*. Here we expanded the bounds around the ideal pharmacophore to include the crystal structure of strychnine, which these authors use as an example of a rigid CNS drug molecule. Although Lloyd and Andrews maintain that the basicity of the nitrogen may vary among CNS drugs, in this query we required that the nitrogen be a basic tertiary amine.

We found 68 structures in CCD that contain this query (i.e., "hits"). Over 40% are well-known CNS-active compounds themselves or are close analogs to CNS-active compounds. These active compounds are listed in Table 1. Some of the structures are shown in Fig. 2*a*. Strychnine, morphine, and mianserin were included in the set of compounds used to derive the pharmacophore. Of the compounds unrelated to these, about 25% are active, a frequency much greater than expected by a random sampling of the CCD. Several hits (besides morphine and strychnine analogs) are rigid or semi-rigid plant alkaloids: latifine, tabersonine, mesembranol, aspidospermidine, etc. For the most part, their biological activities are unknown but are good candidates for testing since many other basic alkaloids have CNS effects. Mesembranol-like alkaloids, for instance, are the basis of the African narcotic drug "Channa" (13).

Mayer *et al.* (14) have found a common set of distances between three key atoms (carbonyl carbon, carbonyl oxygen, carboxylate oxygen) and the enzyme-bound Zn for 28 inhibitors of angiotensin-converting enzyme (ACE). (Many inhibitors are used as antihypertensive drugs.) These distances are shown in Fig. 1*b* for compound **1** (15) in that paper. For our query we redefined this geometry by replacing some distance constraints with angle constraints. This allowed us to express the hydrogen bond geometry around the carbonyl group in a more general way. Also, since the structures in our data base would not contain a Zn atom, we needed to refer to the Zn-binding atom of the inhibitor instead of to the Zn atom itself. We used distance geometry methods (3) to generate 10 conformations of the Zn complex consistent with these distances. We then looked at the range for each distance and angle among all of the conformations. The bounds in our final query include these ranges plus a small interval (0.5 Å for distances, 20 degrees for angles and dihedrals) on either side.

Table 1. Structures in the coordinate data base CCD (see text) that contain the CNS pharmacophore and are known CNS agents or are close analogs to CNS agents

Refcode*	Analogs	Common name	Activity
STRCBH	5	Strychnine	Convulsant
MORPHC	7	Morphine	Analgesic
PRODIL	1	Prodilidine	Analgesic
BOVJIF	1	Preclamol	Analgesic
BIYTAF	2	—	Analgesic
CEHLIK	1	Apomorphinan	Analgesic
MPTHNA	1	—	Analgesic
CIMBAB	1	—	Narcotic antagonist
CIRYAD	1	Benzoctamine	Sedative
BERGUA	1	Ketanserin	Serotonin antagonist
BUCVAW	3	Mianserin	Serotonin antagonist
FANRAN	1	—	Dopamine antagonist
DOBTCL	3	Deoxybutaclamol	Dopamine agonist
MEOBZC†	1	—†	—
DIKSUL‡	1	Fenethazine‡	Antihistamine

*Six-letter identifier of structure in the Cambridge Crystal Database.
†Analog of nefopam, a CNS muscle relaxant and antidepressant.
‡Analog of phenothiazine antidepressants.

For this example, we searched both the CCD and the CLFIL and found 18 and 100 hits, respectively. Selected structures are shown in Fig. 2*b*. In the CCD we found the known ACE inhibitors captopril (MCPRPL), YS-980 (DIVHEV), and enalaprilat (CIYNIH). In the CLFIL we found captopril, an analog of captopril (CL232735) made in-house as part of our antihypertensive program (16), and three structures resembling enalaprilat. These structures are closely related to the compounds Mayer *et al.* used to generate the pharmacophore. It is interesting that a few dipeptides are also found in each data base [e.g., Glu-Val, Glu-Leu, and Glu-Lys in CLFIL and Cys-Gly (CYGNAI) in CCD]. Cheung *et al.* (17) measured the binding of selected dipeptides as possible inhibitors of ACE. None of the peptides mentioned above were tested. However, mercaptopropanoylcysteine, which differs from Cys-Gly only because it lacks the amino nitrogen, bound in the micromolar range. This suggests that further insight might be expected from testing dipeptides of the form Cys-Xaa. Unexpectedly, we found in both data bases a series of dicarboxypyrazines and -pyridines (e.g., APDPZC in CCD), structures in which the essential atoms are locked in the pharmacophore geometry. The superposition of selected hits is shown in Fig. 3.

Our third example is derived from the known structure of a ligand–protein complex. The tryptophan–tryptophan repressor complex has been solved at 2.6-Å resolution (18) and is available as the data set 1WRP in the Brookhaven Protein Databank (19). Fig. 1*c* shows the relationship of amino, carboxylate, and six-membered ring in L-tryptophan in terms of three distances. The equivalent query is shown to the right.

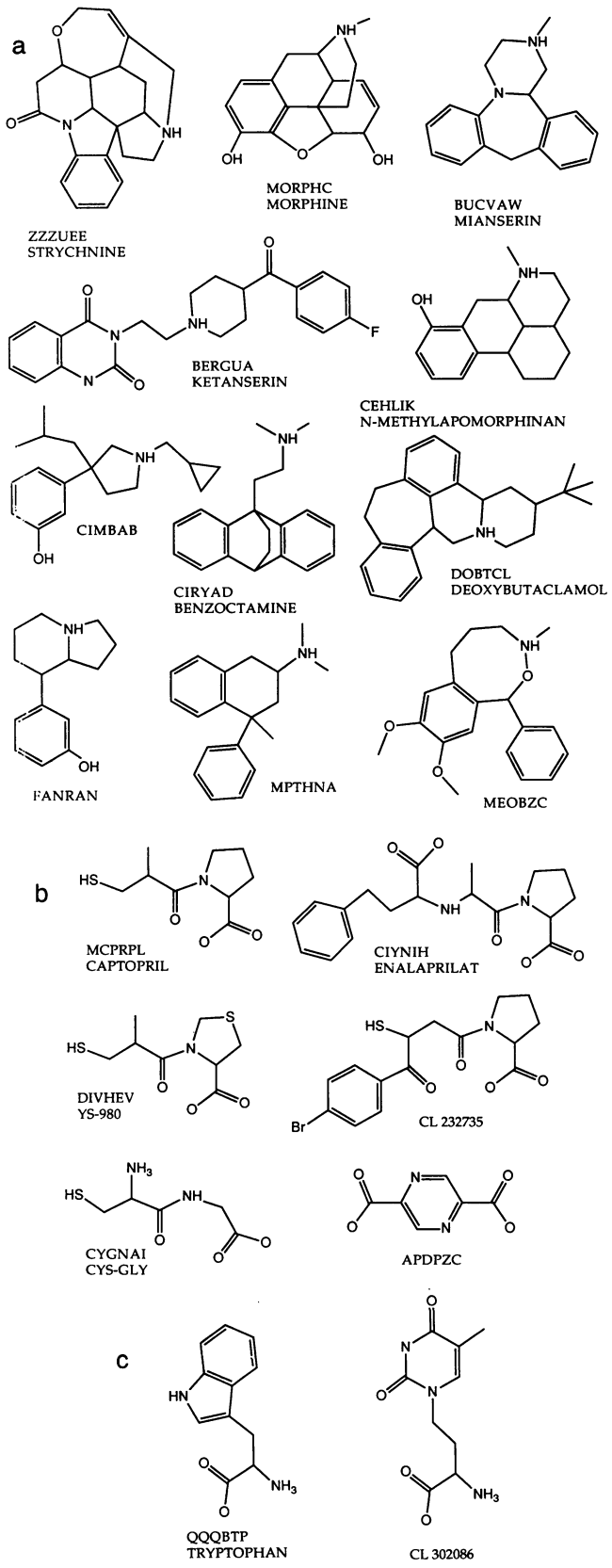


FIG. 2. Selected structures that contain the queries in Fig. 1. In each case, the number of hydrogens on heteroatoms is as stored in the data base (meant to indicate the ionization at physiological pH). Structures from the Cambridge Crystal Database are identified by their six-letter codes and structures from the CL File (Cyanamid Laboratories proprietary data base) by CL numbers. (*a*) CNS pharmacophore. (*b*) ACE inhibitor pharmacophore. (*c*) Tryptophan-repressor pharmacophore.

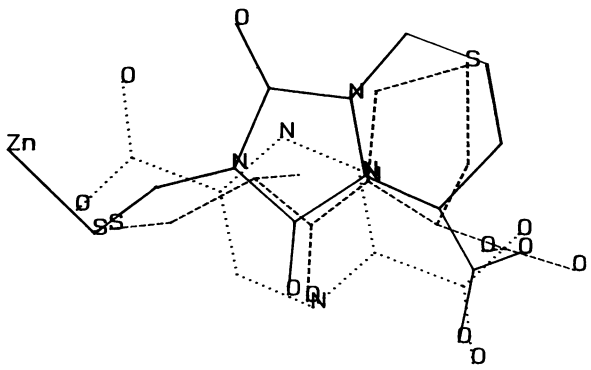


FIG. 3. The superposition of the crystal structure of YS-980 (DIVHEV) (dashed line) and a dicarboxypyrazine (APDPZC) (dotted line) onto one of the conformations (solid line) generated for the structure in Fig. 1*b* by using the distance constraints shown there. The superposition is done at the potential Zn ligand, the neutral hydrogen-bond acceptor, and one of the carboxylate oxygens (atoms 1, 2, and 3 in Fig. 1*b*).

For this example, we want to find structures that contain the pharmacophore and fit into the binding site with little or no steric clash with the protein. Therefore, the volume within 2.4 Å of any of the 122 protein atoms within 7 Å of L-tryptophan is taken as the excluded volume for this query (i.e., the volume in which the center of no atom of a structure may fall).

For this query we found one hit in each data base. These are shown in Fig. 2*c*. These structures in the binding site of the tryptophan repressor are shown in Fig. 4. It is encouraging to note that a search picked the crystal conformation of tryptophan, which is nearly superimposable with the bound conformation in the complex.

DISCUSSION

We have seen that a pharmacophore search can potentially identify: (*i*) compounds in the same chemical class(es) as the compounds from which the pharmacophore was derived (all examples); (*ii*) compounds that contain the pharmacophore, but belong to other chemical classes (many of these will have the appropriate biological activity, as seen retrospectively in the CNS pharmacophore example); and (*iii*) bond "frameworks" in which the important atoms are connected in novel ways or even held rigidly in the desired geometry (as in the ACE-inhibitor example).

Testing the compounds in set *ii* can help us to gain further insight into what molecular features are necessary and/or sufficient for activity and can aid the drug discovery process. In industrial drug discovery, the goal is to find novel "leads" in a particular therapeutic area by testing existing compounds. Set *ii*, being chemically diverse and yet enriched in active compounds relative to a randomly chosen set, could be a particularly fertile source of leads. (For this application, a three-dimensional version of a company's data base of proprietary compounds is especially useful, since this data base is often the sole source of patentable structures and since samples are immediately available for testing.) Fragments in set *iii*, which are unlikely to be found in any type of two-dimensional substructure search, may suggest new directions in the design and synthesis of leads.

It is worth reviewing the limitations to pharmacophore searching. While it is generally accepted that a three-dimensional representation is more "realistic" than a two-dimensional "chemical structure" representation, the complication of conformational flexibility in three dimensions makes generating a pharmacophore model relatively difficult. One needs to have the structure of an appropriate drug–

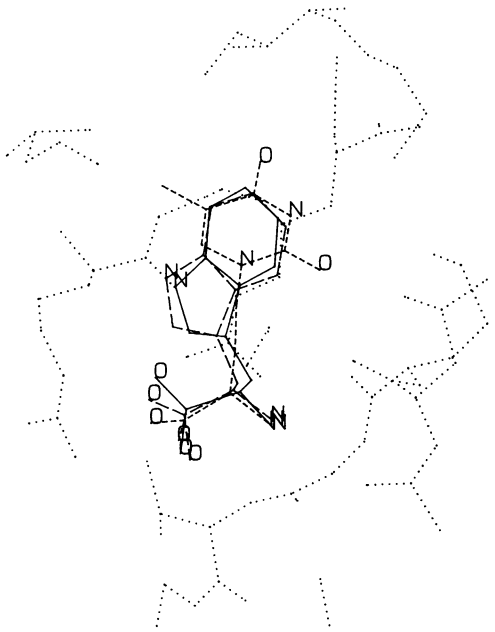


FIG. 4. The crystal structure of tryptophan (QQQBTP) (solid lines) and the CONCORD-generated structure of CL302086 (short dash) docked in the active site of the tryptophan repressor (dotted lines). The docking is done so that the amino nitrogen, the carboxylate carbon, and the centroid of the six-membered ring (atoms 1, 2, and 3 in Fig. 1*c*) of the data base structures are best superimposed onto the corresponding atoms in the bound L-tryptophan (long dash).

receptor complex (for instance, an enzyme and its inhibitor) at the atomic level, or one must deduce the pharmacophore from a set of known drug molecules. In the latter case, one needs a sufficient number of semirigid molecules to define the pharmacophore geometry uniquely. Since the set of molecules used to generate the pharmacophore is necessarily limited in either case, and since many properties of the molecules are not taken into account, there is never a guarantee that the pharmacophore, once generated, is either necessary or sufficient for biological activity.

The second limitation comes from our consideration of only a single conformation per compound: the low-energy crystal conformation in the case of CCD or CONCORD's estimate of a low-energy conformation in the case of CLFIL. These conformations are not necessarily related to the "receptor-bound" conformation represented in a query. Given the number of structures in our data bases (>200,000 for the CLFIL), it is not currently practical in terms of time or disk space to generate and store all of the low-energy conformations or to search through them on a routine basis. The consequence is that if a structure can attain a low-energy conformation consistent with the pharmacophore geometry, but that conformation (or its mirror reflection) is not the one represented in the data base, we will not find that structure. (Of course, in practice, whether or not a given conformation is consistent with a pharmacophore geometry depends on how generously one defines the bounds in the query.) On the other hand, for each structure we do find, it is almost certain that the conformation is energetically accessible.

Fortunately, to be a useful tool, pharmacophore searching need not be exhaustive, only suggestive. For any given pharmacophore, we consistently find compounds that are, in retrospect, known to be active. This leads us to expect that sets of untested compounds derived from such searches will prove enriched in actives. Since in drug discovery the frequency of finding an active compound through random testing is typically a few percent, the enrichment need be only a few fold to be of great significance. Furthermore, only a few novel bond "frameworks" in which important pharmacophore atoms are held in the proper arrangement need to be found to suggest new areas for drug design and synthesis.

1. Marshall, G. R., Barry, C. D., Bosshard, H. E., Dammkoehler, R. A. & Dunn, D. A. (1979) in *Computer-Assisted Drug Design*, ACS Symposium Series 112, eds. Olson, E. C. & Christoffersen, R. E. (Am. Chem. Soc., Washington, DC), pp. 205–226.
2. Crippen, G. M. (1979) *J. Med. Chem.* **22**, 988–997.
3. Sheridan, R. P., Nilakantan, R., Dixon, J. S. & Venkataraghavan, R. (1986) *J. Med. Chem.* **29**, 899–906.
4. Gund, P. (1977) in *Progress in Molecular and Subcellular Biology*, ed. Hahn, F. E. (Springer, New York), Vol. 5, pp. 117–143.
5. Gund, P., Wipke, W. T. & Langridge, R. (1974) in *Proceedings of the International Conference on Computers in Chemical Research and Education* (Elsevier, Amsterdam), Vol. 3, pp. 5/33–38.
6. Allen, F. H., Bellard, S., Brice, M. D., Cartwright, B. A., Doubleday, A., Higgs, H., Hummelink, T., Hummelink-Peters, B. G., Kennard, O., Motherwell, W. D. S., Rodgers, J. R. & Watson, D. G. (1979) *Acta Crystallogr. Sect. B* **35**, 2331–2339.
7. Rusinko, A., Sheridan, R. P., Nilakantan, R., Bauman, N. & Venkataraghavan, R. (1990) *J. Chem. Inf. Comput. Sci.*, in press.
8. Sheridan, R. P., Nilakantan, R., Rusinko, A., Bauman, N., Haraki, K. & Venkataraghan, R. (1990) *J. Chem. Inf. Comput. Sci.*, in press.
9. Martin, Y. C., Danaher, E. B., May, C. S. & Weininger, D. (1988) *J. Comput. Aided Mol. Des.* **2**, 15–29.
10. Jakes, S. E., Watts, N., Willet, P., Bawden, D. & Fisher, J. D. (1987) *J. Mol. Graphics* **5**, 41–56.
11. Ullman, J. R. (1976) *J. Assoc. Comput. Mach.* **23**, 31–42.
12. Lloyd, E. J. & Andrews, P. R. (1986) *J. Med. Chem.* **29**, 453–462.
13. Cordell, G. A. (1981) *Introduction to Alkaloids: A Biogenic Approach* (Wiley, New York), p. 554.
14. Mayer, D., Naylor, C. B., Motoc, I. & Marshall, G. R. (1987) *J. Comput. Aided Mol. Des.* **1**, 3–16.
15. Hassall, C. H., Krohn, A., Moody, C. J. & Thomas, W. A. (1982) *FEBS Lett.* **147**, 175–179.
16. McEnvoy, F. J., Lai, F. M. & Albright, J. D. (1983) *J. Med. Chem.* **26**, 381–393.
17. Cheung, H.-S., Wang, F.-L., Ondetti, M. A., Sabo, S. F. & Cushman, D. W. (1980) *J. Biol. Chem.* **255**, 401–407.
18. Schevitz, R. W., Otwinowski, Z., Joachimiak, A., Lawson, C. L. & Sigler, P. B. (1985) *Nature (London)* **317**, 782–786.
19. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977) *J. Mol. Biol.* **112**, 535–542.