```
3    A(KK)=AA
4    M1=N+1
     BS=B(N)
     DO 5 K=M1,LX
     FF=F(K)
     BB=BS
     BS=B(K)
     F(K)=FF-G*BB
5    B(K)=BB-G*FF
     DO 6 I=1,LA
     K=LA-I+1
6    A(K+1)=-A(K)
     A(1)=1.
     LP=LA+2
     DO 8 I=LP,NP
8    A(I)=0.
     RETURN
     END
```

REFERENCES AND NOTES

(1) Norton, J. P. *An Introduction to Identification*; Academic Press: Orlando, FL, 1986.

(2) Bevington, P. R. *Data Reduction and Error Analysis for the Physical Sciences*; McGraw-Hill: New York, 1969.

(3) Nelder, J. A.; Mead, R. *Comput. J.* 1965, *8*, 308.

(4) Wolfe, M. A. *Numerical Methods for Unconstrained Optimization*; Van Nostrand Reinhold: Workingham, U.K., 1978.

(5) Dutter, R.; Humber, P. J. *J. Stat. Comput. Simul.* **1981**, *13*, 79-114.

(6) Press, H. P.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. *Numerical Recipes*; Cambridge University Press: Cambridge, 1986.

(7) Bickel, P. J. *Ann. Stat.* **1973**, *1*, 597-616.

(8) Jureckovâ, J. *Ann. Stat.* **1977**, *5*, 364-372.

(9) Jaeckel, L. A. *Ann. Math. Stat.* **1972**, *5*, 1449-1458.

(10) Velleman, P. F.; Houglin, D. C. *Applications, Basics, and Computing of Exploratory Data Analysis*; Duxbury Press: Boston, 1981.

(11) Johnstone, I. M.; Velleman, P. F. *J. Am. Stat. Assoc.* **1985**, *80*, 1041-1059.

(12) Adichie, J. N. *Ann. Math. Stat.* **1967**, *38*, 894-904.

(13) Sen, P. K. *J. Am. Stat. Assoc.* **1968**, *63*, 1379-1389.

(14) Rousseeuw, P. J.; Leroy, A. M. *Robust Regression and Outlier Detection*; Wiley: New York, 1987.

(15) Launer, R. L.; Wilkinson, G. N. *Robustness in Statistics*; Academic Press: New York, 1979.

(16) Phillips, G. R.; Eyring, E. M. *Anal. Chem.* **1983**, *55*, 1134-1138.

(17) GAUSS is available from Aptech Systems, Inc., 26250 196th Place South East, Kent, WA 98042 [(206) 631-6679].

(18) Burg, J. P. Paper presented at the NATO Advanced Study Institute on Signal Processing, Enchede, Netherlands, 1968. Also cited in *Nonlinear Methods of Spectral Analysis*; Haykin, S., Ed.; Springer-Verlag: Berlin, 1983; Appendix 6.

Using CONCORD To Construct a Large Database of Three-Dimensional Coordinates from Connection Tables

ANDREW RUSINKO III, ROBERT P. SHERIDAN, RAMASWAMY NILAKANTAN,
KEVIN S. HARAKI, NORMAN BAUMAN, and R. VENKATARAGHAVAN*

Medical Research Division, Lederle Laboratories, American Cyanamid Company,
Pearl River, New York 10965

Received May 5, 1989

The program CONCORD was used to generate a database of three-dimensional (3D) structures from a large (≈265 000 structures) database of connection tables. Additional chemical information was introduced into the 3D database to facilitate the rapid searching for complex 3D substructures such as pharmacophores. Each non-hydrogen atom was characterized by five properties: element type, number of non-hydrogen neighbors, number of π electrons, number of attached hydrogens, and formal charge. For some common functional groups, the number of hydrogens and the formal charge were assigned by considering the most likely ionization state at physiological pH. Dummy atoms were created to represent centroids of flat rings, perpendiculars to those rings, and lone-pair/proton positions. A three-dimensional structural database of 223 988 Cyanamid structures (CL File) was generated. In addition, a subset of the Cambridge Structural Database of experimentally determined molecular geometries was converted into a similar format suitable for 3D substructure search.

INTRODUCTION

Most chemical and pharmaceutical companies maintain large databases of chemical structures containing, among others, structures of their own proprietary compounds. These structures are almost always represented as connection tables (atoms and bonds) and are searchable by topological substructure search systems (e.g., MACCS[1] or the MEDCHEM[2] system). American Cyanamid, like other companies, maintains a collection of samples corresponding to the compounds in the structural database. In conventional screening programs, biologists or chemists may select subsets of structures from the database, so that a small amount of the corresponding sample can be tested in one or more biological assays. This selection can be random (exploring a structurally diverse set for "active" candidates) or guided by some structure-activity model.

Many techniques in the literature attempt to relate the chemical information contained in connection tables (i.e., a two-dimensional structural diagram) to biological activity.[3-6] However, the more sophisticated "graphical" methods of molecular modeling are based on three-dimensional (3D) structures of molecules, and structure-activity rules derived from such methods are naturally expressed in three dimensions. For example, the object of "active analogue" modeling[7-9] is to deduce a pharmacophore, a hypothetical configuration of chemical groups essential for biological activity. One would then like to know which compounds in a proprietary database contain a given pharmacophore. However, since compounds in the database are stored as connection tables and not 3D structures, this type of question is quite difficult to formulate. To properly address 3D questions, one must transform the database of connection tables into 3D molecular structures.

In this paper we will demonstrate how a database of 3D coordinates can be constructed from a large database of connection tables. In the accompanying paper,[10] we discuss

* Author to whom correspondence should be addressed.

0095-2338/89/1629-0251$01.50/0 © 1989 American Chemical Society

a system to search such databases for three-dimensional substructures.

METHODS

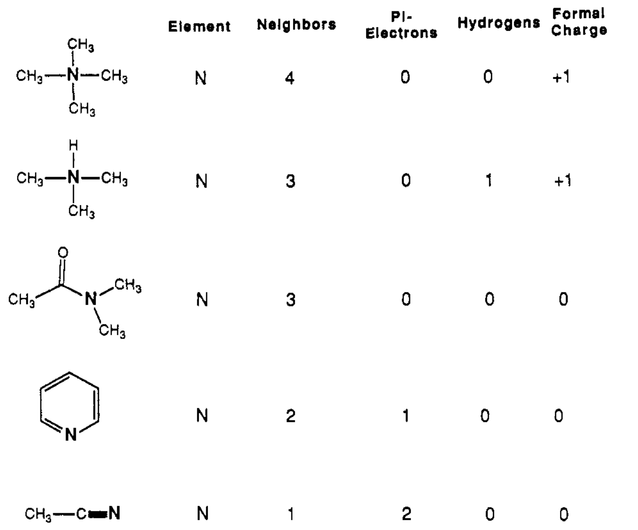
Our strategy for building 3D structural databases can be divided into four steps: 1, generation of 3D coordinates from connection tables; 2, assignment of atom types from connection table information; 3, addition of chemically meaningful "dummy atoms"; 4, efficient storage of the resultant coordinate database. Since steps 2–4 are independent of the source of coordinates, they can be applied as well to databases of preexisting 3D coordinates such as the Cambridge Structural Database.[11]

Generation of Coordinates. Corporate databases of connection tables usually have the following common characteristics: (1) They are large (tens of thousands to *millions* of entries). (2) They are quite diverse. (3) An entry may be multicomponent, i.e., contain more than one molecular fragment. (4) Stereochemical information may be absent, incomplete, or not directly usable. To construct the corresponding database of 3D structures, the method of generating coordinates must be **general** (so that a wide variety of compounds may be treated), **rapid** (so that coordinates may be generated in a reasonable time), and essentially **automatic** (so that little user intervention is required). Three-dimensional structures produced by this technique must also be **accurate** and reliably so.

CONCORD[12,13] is a rule-based program that rapidly and automatically generates a single high-quality approximate 3D structure from its connection table representation. Structures are, by default, produced in a low-energy configuration and conformation that is often comparable to those optimized by molecular mechanics or quantum mechanical methods.[13] These structures are generated in a fraction of the time normally required for conventional sketching and optimization, usually in less than 20 VAX 11/750 CPU s.[13] CONCORD is capable of handling widely diverse classes of compounds containing the "organic" elements (H, C, N, O, F, Si, P, S, Cl, Br, and I) and possessing a maximum individual atomic connectivity of four. Acyclic and mono-, hetero-, and polycyclic structures can be built with a practical upper size limitation of ~300 non-hydrogen atoms. Thus, CONCORD meets all the aforementioned requirements and addresses our primary concerns: speed, accuracy, and the ability to handle diverse chemical classes.

As input, CONCORD (version 2.7) required SMILES (Simplified Molecular Interpretive Line Entry System[14]) strings. We developed a program, TOSMILES, that converts connection tables stored in a MACCS structural database into SMILES strings. Connection tables containing more than one molecule were broken up into component fragments, discarding small counterions (i.e., HCl, SO_4^{2-}, etc.). SMILES strings and 3D coordinates were then generated for each fragment. Second and subsequent fragments were translated 100 Å (in the x direction) from the previously stored fragment.

As noted above, stereochemical information is absent for the vast majority of compounds in our MACCS database. To be consistent in the construction of the initial version of our 3D database, all stereochemical information (even if available) was ignored, even though CONCORD could build structures with the specified stereochemistry. By default, CONCORD positions atoms/groups across a given bond to minimize 1–4 steric interactions. Therefore, configuration about a stereocenter depends on the size of attached ligands relative to previously defined atoms. Ring fusions are more problematic. Although CONCORD has extensive rules for determining the low-energy type (i.e, cis or trans) of ring fusion, occasionally the higher energy ring fusion and not the default type is required to generate the structure correctly. Unfortunately, most of the ring-fusion information was lacking in the MACCS database, and we were forced to accept CONCORD's best guess.



	Element	Neighbors	Pi-Electrons	Hydrogens	Formal Charge
$(CH_3)_4N$	N	4	0	0	+1
$(CH_3)_3NH$	N	3	0	1	+1
$CH_3C(=O)N(CH_3)_2$	N	3	0	0	0
pyridine	N	2	1	0	0
$CH_3—C≡N$	N	1	2	0	0

Figure 1. Example of an atom (nitrogen) in different chemical environments.

Only coordinates for non-hydrogen atoms were retained even though CONCORD generated coordinates for hydrogen atoms as well. This was done primarily to minimize the storage requirements of the 3D database and to keep the structural representation consistent with the MACCS database, which does not include hydrogens. Furthermore, 3D substructural search queries (especially pharmacophores) rarely include explicit hydrogens.

Assignment of Atom Types. The chemical information contained in the CONCORD-generated structures, consisting of 3D coordinates and corresponding element types, was augmented by the addition of information pertaining to each atom's chemical environment implied in the original connection table representation. Thus, each atom was characterized by five descriptors: 1, element type (He–U); 2, number of attached non-hydrogen neighbors (0–8); 3, number of π electrons (0–2); 4, calculated number of attached hydrogens (0–4); 5, formal charge (−1, 0, 1). Examples of atoms represented in this way are shown in Figure 1. Descriptors 1–3 have previously proven very useful in structure–activity studies utilizing topological descriptors.[3,4] Although CONCORD generates structures only for organic compounds, provision was made for the future inclusion of inorganic compounds as well. These structures might include elements with atomic numbers 2 (helium) through 92 (uranium), with each atom having a maximum of eight bonded non-hydrogen neighbors. Atomic numbers between 93 and 96 were reserved for dummy atoms.

The number of hydrogens attached to each atom was computed from the difference between the common valence number and the number of connected non-hydrogen neighbors. However, at physiological pH (≈7.4 in aqueous solutions), some groups become ionized and have more or fewer hydrogens. We felt that the structural databases would be more meaningful pharmacologically if the atomic descriptors reflected the physiological ionization state. The *a priori* prediction of the pK_a of a particular functional group in a molecule is a daunting task. Instead, we identified common functional groups whose ionization state at physiological pH is fairly certain, and assigned to key heteroatoms the number of hydrogens and formal charge depending upon the group's predetermined ionization state. These groups are illustrated in Figure 2. We believe that these cases will cover almost all of the potentially ionizable groups in our databases. Integral formal charges (−1, 0, or +1) were symmetrically assigned

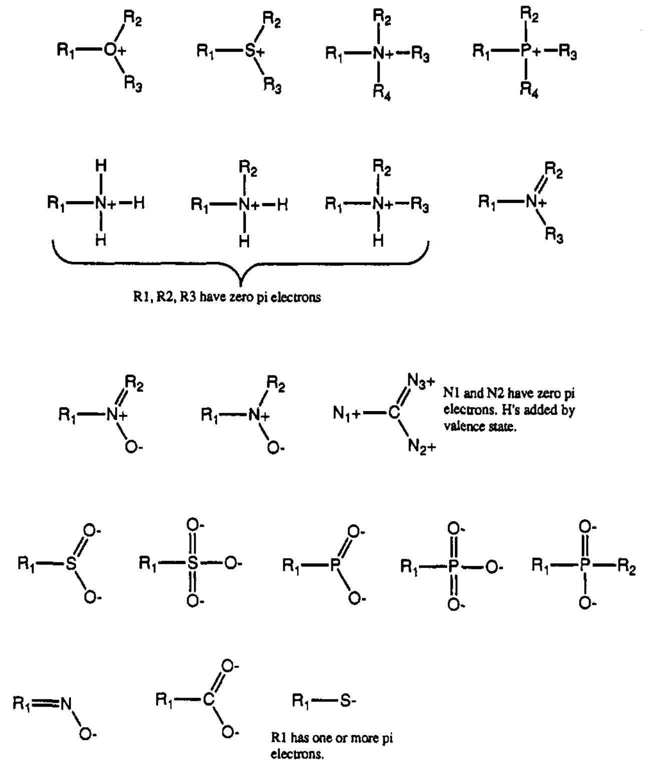

R1, R2, R3 have zero pi electrons
N1 and N2 have zero pi electrons. H's added by valence state.
R1 has one or more pi electrons.

Figure 2. Ionization states of common functional groups.

(e.g., −1 to each oxygen in a carboxylic group) when charge is shared, even though formal charges are not written in this way in standard chemical notation. Charged states that are independent of pH (e.g., quaternary nitrogens) were also assigned using the list of functional groups.

Addition of Dummy Atoms. Three types of dummy atoms created expressly for purposes of three-dimensional substructure searching are depicted in Figure 3. The first type represents centroids of planar 5- and 6-membered rings (designated by pseudoelemental symbols D5 and D6), computed as the mean position of the ring atoms. Ring perpendiculars (DP) were positioned orthogonal to and 0.5 Å above and below each planar ring. Dummy atoms representing lone pairs of electrons (DL) were attached to all heteroatoms and positioned 1.0 Å in the direction of the vector sum of bonds from all neighbors. The chemical meaning of DL depends upon the type of atom to which it is attached. For instance, it can represent the mean lone-pair direction in potential H-bond acceptors or the mean proton direction in potential H-bond donors. The number of non-hydrogen neighbors, the number of π electrons, the number of attached hydrogens, and the formal charge were set to zero for all dummy atoms. A local version of OUTUSR[12] (a user-defined output routine that can be tailored to a specific need and linked directly to CONCORD) was implemented to assign atom types and generate dummy atoms once the structure was built by CONCORD.

Storage of Structures. For each entry in our 3D database, a header record was produced with the identification number, source (CONCORD-generated, X-ray, etc.), conformation number (allowing for the existence of multiple conformers), the total number of non-hydrogen atoms, and the total number of dummy atoms. Next, all information pertaining to each atom in the structure was "packed" for efficient storage and written as an unformated file, indexed by the compound's identification number.

RESULTS

At present, we have completed the generation of 3D structures for the vast majority of Cyanamid compounds. After a preprocessing step that eliminated ≈10 000 organometallic structures or incomplete connection table entries, CONCORD successfully converted 223 988 of the remaining 255 000 connection tables into 3D structures. Generation of the 3D database (including both the translation of the connection tables and the generation of the structures) consumed approximately 40 h of CPU time on our VAX 11/8700. Approximately 70 MB of disk space is required for storage of the database. In addition to the CONCORD-generated 3D database, 30 000 experimentally determined structures from the Cambridge Structural Database, which had available both 3D coordinates and connection table information, were converted into a similar format.

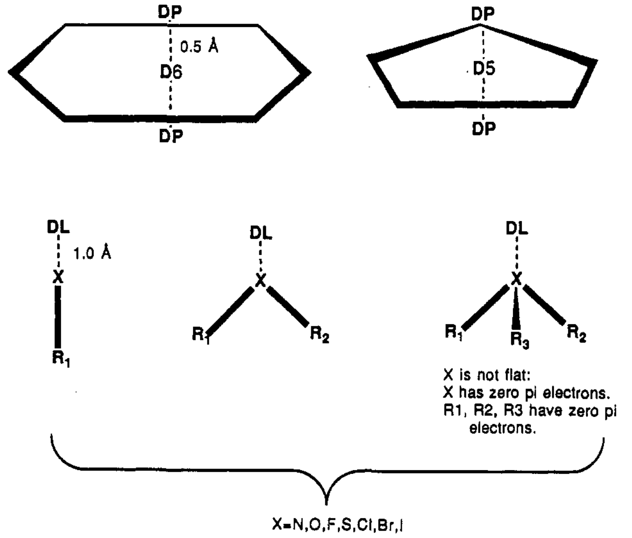


Figure 3. Types of dummy atoms and their relative orientations.

DISCUSSION

In the past, 3D coordinate databases were restricted to small subsets of manually generated structures or to the well-examined Cambridge Structural Database. Gund et al.,[15,16] over 10 years ago, reported one of the first efforts to search 3D databases for pharmacophores. However, these databases were limited to a rather small number of crystal or computer-generated structures. Willett et al.,[17-19] while developing and testing algorithms for rapid 3D substructural searching, performed their studies on a subset of the Cambridge Structural Database. Our intention was to create a large 3D database of proprietary structures that could be easily and rapidly accessed by modern 3D substructure searching techniques[10,17,19] or utilized by other structure–activity methods requiring 3D structures.

It is worth emphasizing the advantages and limitations of generating large three-dimensional databases with CONCORD. Even though CONCORD's structure-building rules are quite extensive and general in nature, one can expect that these rules are not complete, especially considering the structural diversity found in many pharmaceutical/chemical corporate databases. This resulted in a non-negligible fraction of structures (≈10%) that were built incorrectly (and immediately discarded) and another 20% built with CONCORD warnings usually indicative of close van der Waals contacts. Also, several thousand organometallic structures were ignored, since metals are excluded from CONCORD's list of known elements.

Although CONCORD is quite capable of handling stereochemical information, such information is not provided for most entries in our MACCS connection table database. Stereochemical information is important only for compounds that have more than one stereocenter since interatomic distances are obviously the same for each structure and its enantiomer. A survey of a set of randomly selected CL compounds indicated that 73% of the structures had no chiral centers; 24% had one chiral center. For those compounds with the possibility of more than one chiral center, only half had the proper stereochemistry recorded in the connection table. Therefore, we found that, as a first approximation, stereochemistry can be ignored when constructing a large and widely diverse database of 3D structures. However, in some instances, structures requiring stereochemical information and generated without such will yield the wrong epimer or enantiomer. Note that this is still potentially useful information. If, for example, a structure built with the "wrong" stereochemistry serendipitously matches a known pharmacophore, a new lead compound might be suggested by synthetic modification about a stereocenter of an existing compound.

Another potential limitation of CONCORD is that only one low-energy conformation is generated per connection table. That is, the existence of multiple low-energy conformations that might also be available to a molecule is ignored. In general, there is no simple relationship between a single conformation of a molecule (be it solution, X-ray, etc.) and the "receptor-bound" conformation sought when one searches for a pharmacophore. (For example, Jelinski et al.[20] have recently reported that acetylcholine bound to the nicotinic acetylcholine receptor adopts a conformation distinctly different from its conformation in solution or in the crystalline state. The theoretically determined "active" conformation of acetylcholine bound to the muscarinic receptor is also considerably different from that of its crystalline form.)[21,22] On the other hand, since CONCORD builds structures in a low-energy conformation, we are confident that structures retrieved by matching a pharmacophore could indeed exist in such a conformation.

Even if CONCORD had the capability of generating multiple conformations, the storage requirements for just a few low-energy conformations per compound of a large 3D database (hundreds of thousands of structures) would quickly overwhelm many computational facilities. Furthermore, multiple conformations of the same compound would, in effect, add additional 3D structures to a candidate search list and significantly increase the database search time. Also, unless considerable time and effort is expended on each structure, one is never assured that *all* possible low-energy conformers have been located. While quite necessary for rapidly producing alternative solutions, "quick and dirty" methods for randomly generating conformations cannot guarantee that the pharmacologically important conformation might not be overlooked as well. Therefore, we realize that an exhaustive search for all compounds in a large 3D database that could potentially match a given pharmacophore is not currently feasible. However, to be an effective tool in computer-assisted drug design, searches need only be *suggestive*, not exhaustive.[10]

There are alternative methods of converting connection tables to three-dimensional coordinates such as those employing artificial intelligence, WIZARD[23] and AIMB,[24] or distance geometry.[25] Unlike CONCORD, these methods are capable of generating multiple conformations. However, CONCORD remains the most practical method for converting large databases of connection tables. First, CONCORD's structure-building rules were complete enough to generate 3D structures for the overwhelming majority (≈90%) of connection tables. WIZARD and AIMB rely upon previously having "seen" or somehow "learned about" the structure (or more likely portions of it) before they can build a reasonable model. This is not always possible given the quite diverse (and sometimes unprecedented) chemical classes present in corporate databases. Second, to build 3D structures, CONCORD averaged 0.5 CPU per structure on our VAX 11/8700, making the conversion of hundreds of thousands of connection tables in our database possible in less than 2 CPU days. Current distance geometry algorithms cannot achieve the speed required to conveniently process such a large database. Last, user intervention (*human intelligence*), often the ultimate arbitrator for guiding programs to the "best" conformation in other methods, is not required to generate structures with CONCORD.

At this time we know of only one similar conversion of MACCS connection tables to three-dimensional form. Martin et al.[26] made direct use of the MEDCHEM suite of programs.[2] They translated a MACCS database of ≈70 000 structures into SMILES format, converted the SMILES strings to three-dimensional form via CONCORD, and then stored the coordinates as a THOR database.[2]

We are currently implementing three structure-activity methods that will utilize the three-dimensional databases. First, the 3D coordinate database could be used as a source of molecular shapes for shape-searching applications.[27,28] The second and third methods are analogous to one using the more traditional topological databases. One can effectively transform the "atom pair" descriptor[3,4] developed at Lederle into a three dimensional entity. Last, we particularly designed the database so that it could be searched for three-dimensional substructures (for instance, pharmacophores). This application is discussed in detail in the companion paper.[10]

ACKNOWLEDGMENT

We thank David Vanderbrooke for particularly useful discussions concerning indexed files.

REFERENCES AND NOTES

(1) "The Molecular Access System", Molecular Design Limited, San Leandro, CA.

(2) Weininger, D.; Weininger, A.; Leo, A. J. *MedChem Software Manual*, Release 3.52; Medicinal Chemistry Project, Pomona College: Claremont, CA, 1987.

(3) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64-73.

(4) Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological Torsion: A New Molecular Descriptor for SAR Applications. Comparison with Other Descriptors. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 82-85.

(5) Klopman, G. Artificial Intelligence Approach to Structure-Activity Studies. Computer-Automated Structure Evaluation of Biological Activity of Organic Molecules. *J. Am. Chem. Soc.* **1984**, *106*, 7315-7321.

(6) Kier, L. A shape index from molecular graphs. *Quant. Struct.-Act. Relat. Pharmacol., Chem. Biol.* **1985**, *4*, 109-116.

(7) Marshall, G. R.; Barry, C. D.; Bosshard, H. E.; Dammkoehler, R. A.; Dunn, D. A. The Conformational Parameter in Drug Design: The Active Analogue Approach. In *Computer-Assisted Drug Design*; Olson, E. C., Christoffersen, R. E., Eds.; ACS Symposium Series 112; American Chemical Society: Washington, DC, 1979; pp 205-226.

(8) Sheridan, R. P.; Nilakantan, R.; Dixon, J. S.; Venkataraghavan, R. The Ensemble Approach to Distance Geometry: Application to the Nicotinic Pharmacophore. *J. Med. Chem.* **1986**, *29*, 899-906.

(9) Crippen, G. M. Distance Geometry Approach to Rationalizing Binding Data. *J. Med. Chem.* **1979**, *22*, 988-997.

(10) 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.* (following paper in this issue).

(11) 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. The Cambridge crystal data center: computer-based search, retrieval, analysis, and display of information. *Acta Crystallogr. Sect. B: Struct. Crystallogr. Cryst. Chem.* **1979**, *B35*, 2331-2339.

(12) CONCORD, copyright 1987, 1988, The University of Texas at Austin. CONCORD *User's Manual*: TRIPOS Associates: St. Louis, MO., 1988.

(13) Rusinko, A. Tools for Computer-Assisted Drug Design. Ph.D. Thesis, The University of Texas at Austin, Austin, Texas, 1988. Pearlman, R. S. Rapid Generation of High Quality Approximate 3-D Molecular Structures. *Chem. Des. Auto. News* **1987**, *2*(1), 1/5-6. Pearlman, R. S.; Skell, J. M.; Balducci, R.; Rusinko, A. CONCORD: Rapid Generation of High-Quality Molecular Structures. *J. Comput. Chem.* (submitted for publication).

(14) Weininger, D. SMILES, A Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J.*

Chem. Inf. Comput. Sci. **1988**, *28*, 31–36.

(15) Gund, P.; Wipke, W. T.; Langridge, R. Computer searching of a molecular structure file for pharmacophoric patterns. In *Proceedings of the International Conference on Computers in Chemical Research and Education*, Ljubljana; Elsevier: Amsterdam, 1974; Vol. 3, pp 5/33–38.

(16) Gund, P. Three-dimensional pharmacophoric pattern searching. *Prog. Mol. Subcell. Biol.* **1977**, *5*, 117–143.

(17) Jakes, S. E.; Willett, P. Pharmacophoric pattern matching in files of 3-D chemical structures: selection of interatomic distance screens. *J. Mol. Graphics* **1986**, *4*, 12–20.

(18) Jakes, S. E.; Watts, N.; Willett, P.; Bawden, D.; Fisher, J. D. Pharmacophoric pattern matching in files of 3D chemical structures: evaluation of search performance. *J. Mol. Graphics* **1987**, *5*, 41–48.

(19) Brint, A. T.; Willett, P. Pharmacophore pattern matching in files of 3D chemical structures: comparison of geometric searching algorithms. *J. Mol. Graphics* **1987**, *5*, 49–56.

(20) Behling, R. W.; Yamane, T.; Navon, G.; Jelinski, L. W. Conformation of acetylcholine bound to the nicotinic acetylcholine receptor. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 6721–6725.

(21) Schulman, J. M.; Sabio, M. L.; Disch, R. L. Recognition of Cholinergic Agonists by the Muscarinic Receptor. 1. Acetylcholine and Other Agonists with the NCCOCC Backbone. *J. Med. Chem.* **1983**, *26*, 817–23.

(22) Tollenaere, J. P. Muscarinic pharmacophore identification. *Trends Pharmacol. Sci.* **1984**, *5*, 85–86.

(23) Dolata, D. P.; Leach, A. R.; Prout, K. WIZARD: AI in conformational analysis. *J. Comput.-Aided Mol. Des.* **1987**, *1*, 73–85.

(24) Wipke, W. T.; Hahn, M. A. Analogy and Intelligence in Model Building. In *Artifical Intelligence Applications in Chemistry*; Pierce, T. H., Hohne, B. A., Eds.; ACS Symposium Series 306, American Chemical Society: Washington, DC, 1986; pp 136–146. Hahn, M.; Wipke, W. T. Poster session: analogy and intelligence in model building (AIMB). In *Chemical Structures*; Warr, W. A., Ed.; Springer-Verlag, Berlin, 1988; pp 267–278.

(25) Wenger, J. C.; Smith, D. H. Deriving Three-Dimensional Representations of Molecular Structure from Connection Tables Augmented with Configuration Designations Using Distance Geometry. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 29–34.

(26) Martin, Y.; Danaher, E. B.; May, C. S.; Weininger, D. MENTHOR, a database system for the storage and retrieval of three-dimensional molecular structures and associated data searchable by substructural, biologic, physical, or geometric properties. *J. Comput.-Aided Mol. Des.* **1988**, *2*, 15–29.

(27) Sheridan, R. P.; Venkataraghavan, R. Designing novel nicotinic agonists by searching a database of molecular shapes. *J. Comput.-Aided Drug Des.* **1987**, *1*, 243–256.

(28) DesJarlais, R. L.; Sheridan, R. P.; Seibel, G. L.; Dixon, J. S.; Kuntz, I. D.; Venkataraghavan, R. Using Shape Complementarity as an Initial Screen in Designing Ligands for a Receptor Binding Site of Known Three-Dimensional Structure. *J. Med. Chem.* **1988**, *31*, 722–729.

3DSEARCH: A System for Three-Dimensional Substructure Searching

ROBERT P. SHERIDAN, RAMASWAMY NILAKANTAN, ANDREW RUSINKO III,
NORMAN BAUMAN, KEVIN S. HARAKI, and R. VENKATARAGHAVAN*

Medical Research Division, Lederle Laboratories, American Cyanamid Company,
Pearl River, New York 10965

Received May 5, 1989

The system 3DSEARCH is used to search for three-dimensional substructures (for example, pharmacophores) in databases of coordinates. Searches are divided into two parts, a fast prescreen using an inverted key system and a slower atom-by-atom geometric search using the algorithm described by Ullman (*J. Assoc. Comput. Mach.* **1976**, *23*, 31–42). Features to handle angle/dihedral constraints and to take into account "excluded volume" are implemented as part of the geometric search. With this strategy, searches of typically sized queries over large databases (>200 000 entries) take only a few minutes. The speed of the system is demonstrated with a few examples of queries derived from pharmacophores in the literature.

INTRODUCTION

Now that three-dimensional molecular modeling is becoming widely used, "pharmacophore" models for a variety of biological activities are appearing in the literature in increasing numbers. A pharmacophore is a spatial arrangement of chemical groups (usually atoms), common to all active molecules, that is recognized by a single receptor. Pharmacophores can be deduced by a variety of techniques.[1-3] The specifications of the pharmacophores in the literature generally have the following properties: (1) The relationship between the atoms is described by distances and/or angles rather than in terms of bonds. (2) The atoms might be described by chemical property (cation, H-bond donor, etc.) rather than by element type. Also, an atom might be a "dummy", a point that is used to define a geometry, but which takes up no volume (for example, the centroid of a ring).

Often, we want to find which molecules in a set, say, a database of proprietary compounds, contain a particular pharmacophore. To do this we need to have database(s) of coordinates and a method of searching them. In this paper we describe 3DSEARCH, a system to define and search for three-dimensional substructures. The database to be searched may be generated from experimental coordinates (e.g., the Cambridge Structural Database[4]) or from connection tables as discussed in the previous paper.[5]

REPRESENTATION OF ATOM TYPES

For each structure in our database we list a unique identifying integer and the number of non-hydrogen atoms. For each (non-hydrogen) atom are listed the atom type and *xyz* coordinates. As discussed in the previous paper,[5] the atom type consists of five fields: element (He–U); number of non-hydrogen neighbors, i.e., bonded atoms (0–8); the number of π electrons (0–2); the expected number of attached hydrogens (0–4); formal charge (−1, 0, 1). Four types of dummy atoms are used to define geometric points in space. "Element" types D5 and D6 are the centroids of planar 5- and 6-membered rings. DP are along the perpendicular to these rings. DL are connected to each heteroatom and aligned along the sum of vectors from its neighbors.

DEFINITION OF QUERIES

In general, a query is a question formulated to interrogate a database. In our case, a "query" is a three-dimensional substructure consisting of a set of atoms and a description of their spatial relationship. A structure in the database is a "hit" if and only if it contains the query.

Definition of Spatial Relationship between Atoms. The most basic representation of a spatial relationship is as a set of lower and upper bounds to the interatomic distances. A simple example of a structure that contains a query is shown in Figure

0095-2338/89/1629-0255$01.50/0 © 1989 American Chemical Society