Conformational Freedom in 3-D Databases. 1. Techniques

N. W. MURRALL and E. K. DAVIES*

Chemical Design Ltd., Unit 12, 7 West Way, Oxford OX2 0JB, England

Received April 19, 1990

With the advent of databases capable of storing full 3-D information on chemical structures, it is now feasible to search such databases for particular 3-D spatial arrangements of atoms or groups (pharmacophores). This is of particular importance in drug design where the pharmacophore responsible for activity is either already known or can be deduced by standard modeling techniques. Since many of the molecules to be searched are flexible, there is a need to allow for conformational freedom within the database system and to couple it closely to modeling software for further analysis. In this paper, part 1 of a series, we will discuss some techniques that can be used to accomplish these goals as exemplified by the ChemDBS-3D module of the Chem-X modeling suite. Subsequent papers will give details of case studies performed with this system.

INTRODUCTION

In recent years we have seen a rapid growth in the use of three-dimensional databases and search systems for the identification of compounds that match a modeled pharmacophore, in both academic[1] and industrial laboratories.[2-4] The databases searched by these programs are either commercially available[5] or conversions of proprietary 2-D databases,[6] usually built using CONCORD.[7] A typical study would take a pharmacophore derived using standard modeling techniques[8-10] and use this to construct a search query for the database system. The search of the database would reveal any compounds that match the pharmacophore. These results would then require further analysis using modeling and statistical techniques. There is clearly scope for closer integration between the modeling package and the database system to simplify both the process of query generation and that of analyzing the results of the search.

Furthermore, currently available database systems rely on storing a set of individual conformations for flexible molecules, thus increasing both the data-storage requirements and search times for these types of molecules. This approach may also result in compounds being missed since the particular conformation that instills activity may not have been included in the database.

ChemDBS-3D, a module in the Chem-X modeling suite,[11] has been designed to address both the problem of integration with modeling and that of conformational freedom. In this paper we will consider the techniques used to handle conformational freedom and describe how integration with modeling has been achieved. Subsequent papers[12] will discuss a case study giving exact details of the database configuration and the results obtained with the system.

METHODS

The main requirement of a 3-D database searching system to be used by medicinal chemists is to find all compounds that match the modeled pharmacophore. The pharmacophore is normally described in terms of distances between key atoms or sites within the drug molecule, possibly with some substructure requirement. If we take as an example the case of the histamine H_1 pharmacophore,[13] the requirement is for a tautomeric nitrogen system (the substructure requirement) positioned 5.8 and 5.1 Å from a quaternary nitrogen (the geometrical requirement). The process of searching for all conformations of molecules which satisfy these requirements can be considered as three distinct stages: 3-D screening for the pharmacophore pattern, substructure searching, and final generation and matching of individual conformers.

3-Dimensional Screens. All systems that offer high-speed searching, whether 2-D or 3-D, use an initial screen to reduce rapidly the number of compounds under consideration to a very small percentage of the database (typically 1-2%). Screens used by 3-D database search systems indicate that the compound contains two points in space separated by a distance within a particular range.[2,14] The points in space are usually defined by key atom positions within the molecule. This same concept is utilized in ChemDBS-3D, but with two important modifications:

1. The selection of points between which the distances are calculated is not based purely on atom types, but on the chemical properties exhibited by the atom concerned. These active points are termed "centers".
2. A set of representative low-energy conformations (within a user-specified range around the global minimum) from all regions of conformationally accessible space is considered when the screens are generated.

The use of generic atom types avoids the need for the user (or system) to define all individual atom types with the required property and then to perform multiple searches based on all possible combinations of these atom types. Atom types are assigned automatically on the basis of the functional group environment, using substructure search techniques. This enables, for example, nitrogens with a hydrogen substituent to be differentiated from those without. Similarly carbonyl and hydroxyl oxygens can be separated. This functionality is implemented within the Chem-X suite as "level 5 parameterization".

Centers may be defined by the user at database creation time but usually four generic types would be considered:

1. Heteroatoms which can act as H-donors
2. Heteroatoms which can act as H-acceptors
3. Ring centroids
4. Heteroatoms with a formal positive charge

ChemDBS-3D also provides the ability to define specific centers such as nitrogen H-donor or oxygen H-acceptor/lone pair. However, the use of these could bias the results since it is frequently the property of the atom that is the important factor not the actual atomic type.[15]

These centers are automatically determined from a structure either built within Chem-X or read into the Chem-X system, thus enabling the medicinal chemist to enter a query without specialist knowledge of complex 3-D objects or how the database system works.

By applying these concepts to the histamine H_1 example previously discussed, we see that it consists of a 1-3 H-donor-H-acceptor system at a separation of 2.2 Å, with an H-donor at 5.1 and 5.8 Å from those two centers.

* Author to whom correspondence should be addressed.

0095-2338/90/1630-0312$02.50/0

The presence of a particular separation is indicated by setting a screen that represents the presence of that particular feature within a range encompassing the distance. The screens are arranged in sets, where each set contains the screens relevant to a particular type of interaction, e.g., donor–donor or acceptor–donor interactions. Each screen within the set, termed a "bin", is used to indicate an interaction within a particular distance range, e.g., 5.0–5.5 Å. Thus a donor–donor separation of 5.8 Å would set the screen for donor–donor separations between 5.5 and 6.0 Å. The distance ranges used in the screens are not distributed linearly over all accessible distances, and in fact should be distributed such that each distance range is evenly populated.[16] This may be calculated empirically from the actual database to be used,[14] or more conveniently, ranges may be derived from the expression

$$\text{bin number} = X \times \arctan\,[(D - 3.0)/2] + Y$$

where X and Y are constants and D is the interaction distance since this has been found to provide quite a reasonable representation of the distribution.[2]

While assigning distance ranges to the bins, two other factors should be taken into consideration: the ability to screen for functional groups and an upper bound to the distances considered. Common functional groups, particularly carboxylic acids and other 1-3 substituted systems, 1-4 hetero rings, and substituted rings such as phenols, have key distances of ca. 2.20, 2.80, and 2.75 Å, respectively, which should be identified by specific bins. Willett has shown that above 15 Å there is little point in subdividing into ranges[14] since such distances occur relatively infrequently.

The implementation in ChemDBS-3D uses one word (32 bits) per screen set, with one bit representing a bin. For maximum screen search speed it was also necessary to allow use of the FORTRAN IAND instruction for any combination of bits. This allows 31 discrete distance ranges in the system since the sign bit is not usable. Thus, for example, small distances in the range 1.7–3.0 Å would use an increment in bin size of 0.1 Å, all distances below 1.7 Å set a single bin, and all distances above 15 Å could also set a single bin. A typical distribution would therefore be

distance < 1.7	bin 1
1.7 < distance < 3.0	bins 2–14 step 0.1 Å
3.0 < distance < 7.0	bins 15–22 step 0.5 Å
7.0 < distance < 15.0	bins 23–30 step 1.0 Å
distance > 15.0	bin 31

although this may be defined by the user at database configuration time. The bin sizes for distances above 3.0 Å reflect the uncertainty in typical pharmacophore patterns. This method for assigning the bin sizes thus represents a compromise between identification of well-defined functional groups and screening for specific pharmacophore distances.

Formula Screen. A formula screen is also used in the system. This, too, is based on the center definitions and was designed to screen out those compounds that do not possess the number of centers of each type required by the query. This is combined with atom-type information to give a more specific key. Thus the number of nitrogen H-donors, nitrogen H-acceptors, oxygen H-donors, oxygen H-acceptors, etc. is stored for each compound as well as the total number of generic H-acceptors, H-donors, centroids, etc. The count is stored by setting one, two, or more bits, up to a configurable maximum, for each type of center found in the molecule. Thus if three bits were allocated to storing the presence of generic H-donors, these would represent the presence of one, two, and more than two H-donors, respectively. Use of one bit would indicate the presence or absence of a particular feature. Typical features included in the formula key and the abbreviation used (with the number of bits allocated to the count in parentheses) would be

1. generic H-donors, D's (3)
2. generic H-acceptors, A's (3)
3. charge centers, + (2)
4. specific nitrogen H-donors, ND (3)
5. specific nitrogen H-acceptors, NA (3)
6. specific oxygen H-donors, OD (3)
7. specific oxygen H-acceptors, OA (3)
8. six-membered (hetero)aromatic rings, 6 (1)
9. five-membered (hetero)aromatic rings, 5 (1)

The formula key for histamine with two H-donors (nitrogen), one H-acceptor (nitrogen), a charge center, and a five-membered heteroaromatic ring based on the above scheme would then be

```
                    |5|6|OA |OD |NA |ND |+ |A's|D's
0000000000|1|0|000|000|001|011|01|001|011
```

Since the presence of many functional groups can be derived from the 3-D distance keys, there is no need to specifically include these features within the formula screen. The main exceptions are aromatic rings, which may not have a readily identifiable pattern of centers.

Generating the Screens. The process by which these screens are calculated for a molecule is described by the following steps:

1. Locate the centers of the molecule and determine their types.
2. Identify rotatable bonds within the structure.
3. Perform conformational analysis on the molecule and at each step calculate the relevant set of screens for that static conformation.
4. For all low-energy conformations (those that fall within the predefined range of the global minimum) combine all screens of the same type (e.g., donor–donor distances) into one screen using a logical OR.
5. Store the combined screens, formula key, and energy of the minimum energy conformer in the database.

The identification of rotatable bonds finds only acyclic bonds and assigns the number of steps to be used on the bond type. Double bonds would use 180° steps, those between sp^2 and sp^3 atoms 60° steps, and those between two sp^3 atoms 120°. Chem-X rules[17] would also usually be applied during the analysis.

For histamine the set of screens produced using the above distribution of bins and using 10 low-energy conformers generated by Chem-X rule-based[17] analysis are

```
Donor-Donor      000 0000 0000 1010 0000 0000 0000 0000
Donor-Acceptor   000 0000 0000 0001 1010 0000 0100 0000
```

The 1-3 substituted imidazole ring is clearly indicated by the setting of bin 7 (bins are counted from the right) in the donor–acceptor screen. The various conformations are indicated by the setting of bins 18 and 20 in the donor–donor screen and bins 14, 16, and 17 in the donor–acceptor screen. All screens are stored as bit maps in files that can rapidly be accessed by the system using the VAX/VMS utility SYS$CRMPSC, thus optimizing search speed. These files form a natural extension to the three-dimensional modeling database ChemDBS-1 (part of Chem-X), which is used to store coordinates, bonding, atom types, and associated data.

These screens allow rapid searching of all accessible conformations at a rate that is independent of the number of conformations stored.

Screen Searching. A query for ChemDBS-3D is formulated from a structure, or fragments, modeled using the Chem-X system. This may be either sketched in with standard 3-D building techniques or read in from a disk file. The query is then automatically perceived directly from this structure, with

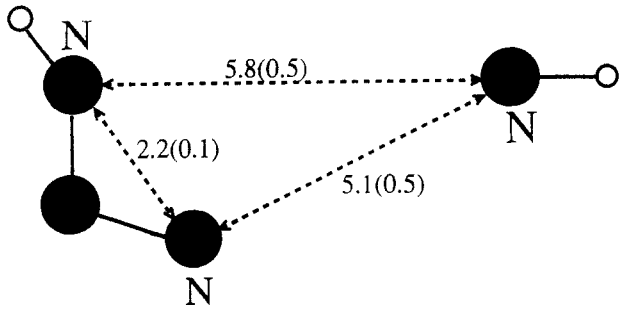

Figure 1. Histamine H_1 pharmacophore query.

no further intervention by the chemist being required.

As in screen generation, the centers and their types are identified automatically and the distances between them calculated. Tolerances are applied to these distances, typically 0.1 Å between bonded centers, centers bonded to a common atom, or between a ring centroid and a center used in defining that ring, and 0.5 Å for all other interactions. However, these tolerances may be changed by the user for any pair of interacting centers.

Thus, if the chemist were searching for compounds analogous to the histamine H_1 active conformation (all trans), the query would be constructed from two fragments, a tautomeric nitrogen system and an N–H group as shown in Figure 1. It contains an H-donor–H-donor interaction of 5.8 ± 0.5 Å and two H-donor–H-acceptor interactions, one of 5.1 ± 0.5 Å and the other of 2.2 ± 0.1 Å. These interactions generate the following search screens:

```
Donor-Donor     0000 0000 0001 1100 0000 0000 0000 0000
Donor-Acceptor  0000 0000 0000 1110 0000 0000 0000 0000
Donor-Donor     0000 0000 0000 0000 0000 0000 1110 0000
```

A formula key is also constructed which indicates the need for at least two generic H-donors and one H-acceptor, two specific nitrogen H-donors and one nitrogen H-acceptor.

0000 0000 0000 0000 0000 1011 0000 1011

The screening process then takes the following form. For each record in the database:

1. The formula key is checked by using a logical AND. A result equal to the search formula key indicates that the required features are present.
2. Each interaction key is checked against the corresponding key type in the database using a logical AND. A nonzero result indicates at least one bit common with the query, and hence the screen criterion is satisfied for that interaction.
3. If the above criterion is satisfied for all interactions, the database compound is stored in the answer set.

The answer sets are stored as an array of bit-masks (one bit-mask per compound) with each bit indicating the presence of that compound in the corresponding answer set.

It should be noted that the screen search only excludes those compounds that cannot match the query. The resulting compounds do not necessarily all match the query since there probably will not be a perfect match between the range corresponding to the bins used in the screening and that specified by the query. For example, a compound with an H-donor–H-donor distance of 6.45 Å would pass the screen criterion since this corresponds to the bin for 6.0–6.5 Å which falls in the query range, but it does not match the query exactly. Exact geometry matching is performed later in the search sequence.

Substructure Search. The substructure search is an important, but not essential, part of the system. It fulfills two main roles:

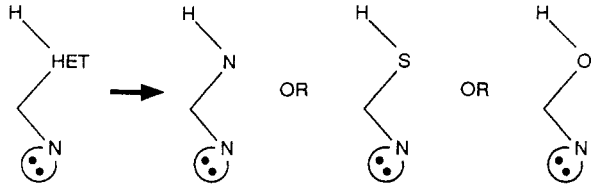

Figure 2. Effect of using generic atom types on the 1-3 nitrogen system.

Table I. Average Search Rates of Component Parts of ChemDBS-3D (Compounds per CPU Second, VAX 8600)

search method	rate
screen	100 000
substructure	>10
field	150
final matching	3[a]

[a] See text.

1. to ensure that any substructure requirement is met by the candidate structures
2. to identify all possible candidates for centers corresponding to those in the query

The search is based on the connectivity and atom types of the structure defined as the query within the Chem-X system. This again enables the non expert database user to perform the search quickly and easily. It is also normally restricted to the results of a 3-D screen search or a field search (see below). In addition to the modeled atom types, additional allowed atom types and wild cards can be specified for any atom in the query. Generic atom types for hetero atoms (HET), halogens (HAL), the alkyl atoms C and H (R), non-hydrogen (NONH), or any atom (AA) allow common generic queries to be easily specified. Thus for the 1-3 tautomeric system in histamine the nitrogen donor may be defined as the generic type HET, implying any of the elemental types N, S, or O (see Figure 2).

The algorithm used employs standard subgraph isomorphism techniques with tree search and backtracking,[18] searching by depth first, and is used to identify all possible matches of the query within the candidate structure. The ability to handle several disconnected fragments was quickly identified as an important requirement of such an algorithm.

The compounds that match the query are stored in the answer set as for screen searching.

Although not optimized for pure substructure searching and being limited by disk access for all the three-dimensional data, search speeds compare favorably to other commercially available systems,[19,20] at more than 10 compounds per CPU second (see Table I).

Generation of Final Conformers. The final part of the search is the generation of the various molecular conformations that match the original query. This is accomplished by once again performing conformational analysis on the candidate structures and determining all those conformations which match the pattern of centers defined by the query. The conformations are regenerated since this is quicker than reading all the required data from disk, and it also saves on disk storage space since only one conformation need be stored. It is usually performed immediately after the substructure search since the data about possible candidate atoms corresponding to the query centers are directly available. The algorithm used to determine if the query pattern is satisfied is that of Ullman[21] as described by Brint and Willett.[22] The basic sequence of operations is as follows:

1. Analyze the molecule for rotatable bonds and assign as for the key generation.
2. Find all pairs of interacting centers that were uniquely defined in the substructure search. Enter the

interaction distance as a constraint on the conformations to be generated.

3. Use the substructure data to initialize the matching matrix M_0 of the Ullman algorithm so that only those possible candidates already identified are considered as centers.

4. Generate the conformations. For each conformation not excluded by the distance constraints:

a. Determine whether it is a low-energy conformation using the same criterion as applied during key generation.

b. If so, determine whether the pattern of centers matches that of the query by performing the subgraph isomorphism test using the Ullman algorithm.

c. If so, rename the atoms in the structure so that the atom-naming scheme corresponds to that of the query molecule.

d. Orient the structure so that the corresponding centers are superimposed on one another by least-squares fit.[23]

e. Write out the conformation to a new results database.

f. Write the compound identifier to the answer set.

With rigid molecules the structure read from the database is analyzed to determine whether it fits the pharmacophoric pattern. If no substructure data is available, the matching matrix M_0 is initialized using the center types only, and no criteria can be set for the generation of the conformations, thus increasing the search time.

This search procedure generates two sets of answers:

1. The non conformationally dependent answer set.

This is a list of the compounds that have at least one low-energy conformation that matches the query. It is associated with the original database, and therefore the compounds are not renamed, reoriented, or aligned with the query and may not be the actual conformation that matched the query. Further studies that do not depend on the actual 3-D structure may be performed with this answer set, e.g., pure activity studies.

2. The conformationally dependent answer set.

This is the database produced during conformation regeneration and contains all the molecular conformations that match the query. These are renamed and reoriented to match the query. This database would be used for further modeling studies where the 3-D structure of the potential candidate is important.

The speed of this part of the system is dependent to a large extent on the flexibility of the molecules and is also limited by the disk access required to produce the results database. Energy and pattern matching alone, without creating the results answer set, can be performed at the rate of 100 conformations per CPU second. For example, in the case of dopamine analogues with three flexible bonds about 50 conformations require energy testing and pattern matching, of which six are accepted. Averaged timings indicate that 2.2 CPU s are required to process each molecule or about 3 hits per CPU second. This compares very favorably with other reported rates of 0.3–7 and <0.4 hits per CPU second.[2,4] It must however be borne in mind that these studies used different CPU's and also that neither of these systems produces a local database of oriented molecules.

Field Searching. In addition to the search capabilities based on 3-D structure and connectivity, ChemDBS-3D also offers the user the ability to search by molecular property. These may be parameters calculated with standard modeling techniques or experimentally observed values such as biological activity.

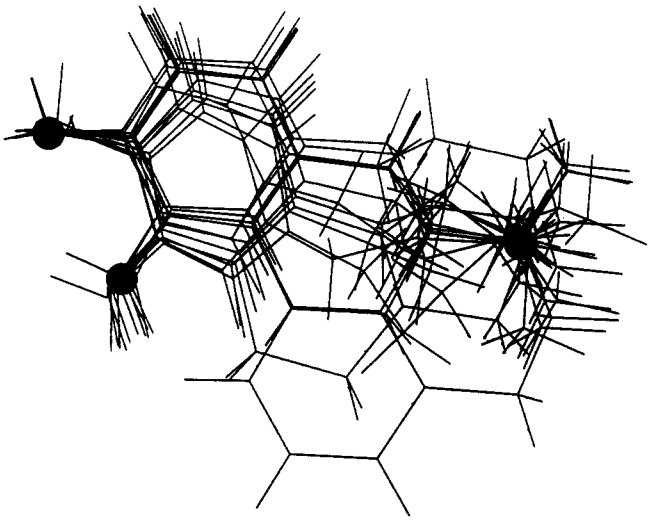

Figure 3. Superpositioned conformations resulting from a search for dopamine analogues.

The properties are stored in any of the 104 fields available in the associated ChemDBS-1 database, and a range of acceptable values for the property can be specified by the user. A typical use of this kind of search would be to consider only those compounds for which a sufficient sample is available for physical testing or those with pK_a or log P values within certain ranges. Another example would be to search on complex 3-D geometry variables such as angles between planes or dipole alignment. These complex geometry calculations would be performed automatically by using the ChemStat[24] module of Chem-X before the field search is initiated.

Processing the Results. Several features have been incorporated into the system in order to facilitate the integration of the database search with modeling activities.

We have already noted that the query is specified using the modeling system rather than specialist database software. This makes the system more accessible to our target user, the medicinal chemist at the bench, who may not be familiar with the mathematical concepts required by other systems.[19]

The automatic superpositioning and renaming of the resultant structures in the conformationally dependent answer set greatly enhance the user's ability to proceed to further modeling studies (see Figure 3). These may involve spatial analysis for the active analogue approach[25] or complex geometry or quantum mechanical property calculations[26] which could then be subjected to rigorous statistical analysis using ChemStat.[24]

Similarly the non conformationally dependent answer set may be studied further using ChemStat if desired.

The full display facilities of the modeling system are also available to the user. These enable a wide variety of display styles to be used, allowing very rapid visual comparisons to be made, further enhanced by the common orientation used in the results database. High-performance graphics workstations allow the use of real-time manipulations and animation techniques, by which the user can rapidly cycle through all the molecular conformations generated.

In line with current trends in open architecture, answer sets may be written to and read from simple ASCII files. This allows searches on data not present in the ChemDBS-1 database to be performed and the results brought back into Chem-X for further analysis. A typical case would be where biological activity data is stored on a relational database system such as ORACLE.[27] The answer set of the 3-D screen search could be exported to ORACLE and searched for those compounds with specified activity and then the results imported back into Chem-X for further substructure analysis and

conformation generation. Answer sets may also be combined by using logical AND, NOT, and OR operations.

CONCLUSIONS

The system described above satisfies the prime requirements of a 3-D search system in that it

1. Is able to search effectively all accessible conformations of flexible molecules within realistic time scales.
2. Integrates both query construction and subsequent results processing with a modeling system.

It should be noted that no step in the process is dependent on another step having been taken previously, the suggested sequence of events being given for efficiency only.

As a consequence of the methods used, screen search times are independent of the number of conformations handled, as are the database disk storage requirements.

Subsequent papers in this series will give details of actual database configuration parameters and describe a typical case study.

REFERENCES

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

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

(3) Martin, Y. C.; 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.

(4) Van Drie, J. H.; Weininger, D.; Martin, Y. C. ALADDIN: An integrated tool for computer-assisted molecular design and pharmacophore recognition from geometric, steric, and substructural searching of three-dimensional molecular structures. *J. Comput.-Aided Mol. Des.* **1989**, *3*, 225-251.

(5) Allen, F. H.; Bellard, S.; Brice, M. D.; Cartwright, B. A.; Doubleday, A.; Higgs, H.; Hummelick, T.; Hummelick-Peters, B. G.; Kennard, O.; Motherwell, W. D. S.; Rodgers, J. R.; Watson, D. G. The Cambridge Crystallographic Data Center: Computer-Based Search Retrieval, Analysis and Display of Information. *Acta Crystallogr.* **1979**, *B35*, 2331-2339.

(6) Rusinko, A., III; Sheridan, R. P.; Ramaswamy, N.; Haraki, K. S.; Bauman, N.; Venkataraghavan, R. Using CONCORD to construct a Large Database of Three-Dimensional Coordinates from Connection Tables. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 251-254.

(7) Rusinko, A., III; Skell, J. M.; Balducci, R.; McGarity, C. M.; Pearlman, R. S. CONCORD: *A program for the rapid generation of high quality approximate 3-dimensional molecular structure*; The University of Texas: Austin; and Tripos Associates: St. Louis, MO, 1988.

(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) Mayer, D.; Naylor, C. B.; Motoc, I.; Marshall, G. R. A unique geometry of the active site of angiotensin-converting enzyme consistent with the structure-activity studies. *J. Comput.-Aided Mol. Des.* **1987**, *1*, 3-16.

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

(11) Chem-X molecular modeling software, developed and distributed by Chemical Design Ltd., Oxford, England, 1990.

(12) Murrall, N. W.; Davies, E. K. In preparation.

(13) Ganellin, C. R. Chemistry and Structure-Activity Relationships of Drugs Acting at Histamine Receptors. In *Pharmacology of Histamine Receptors*; Ganellin, C. R., Parsons, M. E., Eds.; Wright-PSG: London, 1982; pp 35-37.

(14) 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.

(15) Ganellin, C. R. Chemistry and Structure-Activity Relationships of Drugs Acting at Histamine Receptors. In *Pharmacology of Histamine Receptors*; Ganellin, C. R., Parsons, M. E., Eds.; Wright-PSG: London, 1982; pp 21 and 77.

(16) Ash, J. E.; Chubb, P. A.; Ward, S. E.; Welford, S. M.; Willett, P. *Communication, Storage and Retrieval of Chemical Information*; Ellis Horwood Ltd.: Chichester, U.K., 1985; pp 160-167.

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

(18) Tarjan, R. E. Graph algorithms for chemical computation. *ACS Symp. Ser.* **1977**, *No. 46*, 1-19.

(19) MACCS—Molecular ACCess System, developed and distributed by Molecular Design Ltd., San Leandro, CA, 1989.

(20) OSAC—Organic Structures Accessed by Computer, developed and distributed by ORAC Ltd., Leeds, England, 1989.

(21) Ullman, J. R. An algorithm for subgraph isomorphism. *J. Assoc. Comput. Mach.* **1976**, *16*, 31-42.

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

(23) Ferro, D. R.; Herrmans, J. A different best rigid-body molecular fit routine. *Acta Crystallogr.* **1977**, *A33*, 345-347.

(24) Perry, N. C.; Davies, E. K. The use of 3D modeling Databases for Identifying Structure Activity Relationships. In *QSAR: Quantitative Structure-Activity Relationships in Drug Design*; Fauchere, J. L., Ed.; Alan R. Liss Inc.: New York, 1989; pp 189-193.

(25) Marshall, G. R.; Barry, C. D.; Bossard, H. E.; Dammkoehler, R. A.; Dunn, D. A. 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.

(26) Young, R. C.; Durant, G. J.; Emmett, J. C.; Ganellin, C. R.; Graham, M. J.; Mitchell, R. C.; Prain, H. D.; Roantree, M. L. Dipole Moment in Relation to H_2 Receptor Histamine Antagonist Activity for Cimetidine Analogues. *J. Med. Chem.* **1986**, *29*, 44-49.

(27) ORACLE: Relational Database Management System, distributed by ORACLE U.K. Ltd., Richmond, Surrey, England, 1989.

Automated Conformational Analysis and Structure Generation: Algorithms for Molecular Perception

ANDREW R. LEACH,* DANIEL P. DOLATA,‡ and KEITH PROUT

Chemical Crystallography Laboratory, University of Oxford, 9 Parks Road, Oxford OX1 3PD, U.K.

Received May 2, 1990

Many methodologies for performing automated conformational analysis require some means of "perceiving" a molecule to determine features of interest. Algorithms for finding rings, bond orders, and stereocenters and detecting the presence of substructural fragments have been developed. These algorithms are described, emphasizing their importance in conformational analysis.

INTRODUCTION

WIZARD and COBRA are two programs which use artificial intelligence techniques to construct low-energy conformations of molecules. One of the major objectives of this project was to develop a means by which low-energy conformations of a molecule could be rapidly generated, starting from a simple "two-dimensional" representation (i.e., connectivity and atom types). In this paper some of the algorithms that these programs use to "perceive" a molecule are described. First, however, a brief description of the method these programs use to perform a conformational analysis is given (further details can be found elsewhere[1]).

Starting from a definition of the molecule (which can be specified either via a datafile obtained from a molecular

* Address correspondence to this author at his present address: Computer Graphics Laboratory, School of Pharmacy, University of California, San Francisco, CA 94143-0446.

‡ Present address: Dept. of Chemistry, University of Arizona, Tucson, AZ 85721.

0095-2338/90/1630-0316$02.50/0