Experimental System for Similarity and 3D Searching of CAS Registry Substances. 1. 3D Substructure Searching†

William Fisanick,* Kevin P. Cross, J. Christopher Forman, and Andrew Rusinko III

Research Unit, Chemical Abstracts Service, 2540 Olentangy River Road, P.O. Box 3012, Columbus, Ohio 43210

Received February 4, 1993

Chemical Abstracts Service (CAS) is developing an experimental system for similarity and 3D searching on CAS Registry substances. The purpose of this system is to obtain user input on desirable capabilities and data content for such searching. Currently, the system supports 3D exact, substructure, and superstructure searching. The 3D coordinates for the system's databases were generated via the CONCORD program. These databases include a CAS 3D structure templates (CAST-3D) subset of over 365 000 substances with limited conformational flexibility. The experimental system utilizes a client–server architecture using client workstations for query framing and display and a single search engine compute server. Search levels include a screen step followed by atom-by-atom search (using a modified Ullman subgraph isomorphism algorithm), and, where appropriate, a geometric superimposition of the query and answer file substance. Two logical 3D substructure query types are supported: a general query, typically used for pharmacophore pattern matching, and a framework query, typically used in locating synthetic precursors that lead to a desired geometric orientation of substituents. Novel screens based on atom triangle and tetrangle distances as well as global and localized flexibility indices provide for effective and efficient screening. Also, user parameters can specify approximate local and global conformational flexibility characteristics for the matching file substances. This paper describes the features and capabilities that are currently available in the experimental system along with an illustrative application scenario.

I. INTRODUCTION

The scope of chemical information handling has expanded in recent years to include the storage, search, and retrieval of three-dimensional (3D) structure representations in addition to the traditional two-dimensional (2D) representations of chemical substances. The focus of 3D structure search and retrieval has been substructure (inclusion-match) searching, primarily to locate pharmacophoric patterns of atoms. Research activity on 3D substructure searching was initiated in the mid 1970s and early 1980s.[1-3] This activity increased considerably in the late 1980s[4-22] due in part to the availability of the CONCORD program[23,24] which can rapidly generate a set of high quality, approximate 3D coordinates from the 2D structures of organic substances.

Since the early 1960s, Chemical Abstracts Service (CAS) has experimented with and developed substructure search capabilities based on 2D structure, nomenclature, and Markush structure representations of chemical substances.[25-34] Recently, CAS has been exploring several additional capabilities which could expand our scope of handling of chemical substances. This includes research activities on 3D substructure searching and similarity (fuzzy-match) searching on 2D structures, 3D structures, and molecular property data.[35] CAS currently has 3D coordinates for over 4.6 million Registry substances. These coordinates were generated by the CONCORD program and currently are available only on a retrieval basis via the Scientific and Technical Information Network (STN). One objective of on-going research is to ascertain the feasibility of directly searching large files of 3D structures and related molecular properties. The molecular property data have been computer-generated from mostly 3D structure data using computational chemistry programs.

These new capabilities are being incorporated into an experimental system designed to obtain user input on their desirability, including the data content of the databases. Other objectives of the CAS similarity and 3D experimental search system (SIM3D/ESS) are to ascertain the feasibility and desirability of scaling the system to handle several million substances, to provide the basis for an experimental search service which is expected to begin in early 1993, and to help test a client–server interface model between CAS and other vendor search systems, such as ISIS/MACCS-3D[36] and SYBYL/3DB UNITY,[37] via a wide area network.

Currently, 3D exact, substructure, and superstructure searches have been implemented in SIM3D/ESS. While a general similarity search functionality has not yet been added to the system, a considerable amount of experimentation on fuzzy-match techniques for 3D structure and molecular property data has been performed. The detection of significant size and shape similarity based on generic 3D triangle fragments and significant isosteric or chemical similarity based on global molecular property features has been previously reported.[35] This paper describes the features and capabilities that are currently available in the experimental system along with an illustrative application scenario.

II. SYSTEM ARCHITECTURE

The SIM3D/ESS consists of two main components: the query framing & display facility (QDF) "client" and a search facility "server". A QDF client is an interface for framing 3D substructure queries and displaying and manipulating 3D answer structures. It executes on a SUN SPARCstation-2 or a Silicon Graphics IRIS workstation and is based on Tripos Associates SYBYL molecular modeling software (versions 5.5 and 6.0).[38] Thus, the extensive 3D structure manipulation features of the SYBYL software package are available to a QDF. The QDF command set is written in SPL (SYBYL

† Presented in part at the 204th ACS National Meeting held in Washington, D.C., Aug 23–28, 1992.

0095-2338/93/1633-0548$04.00/0 © 1993 American Chemical Society

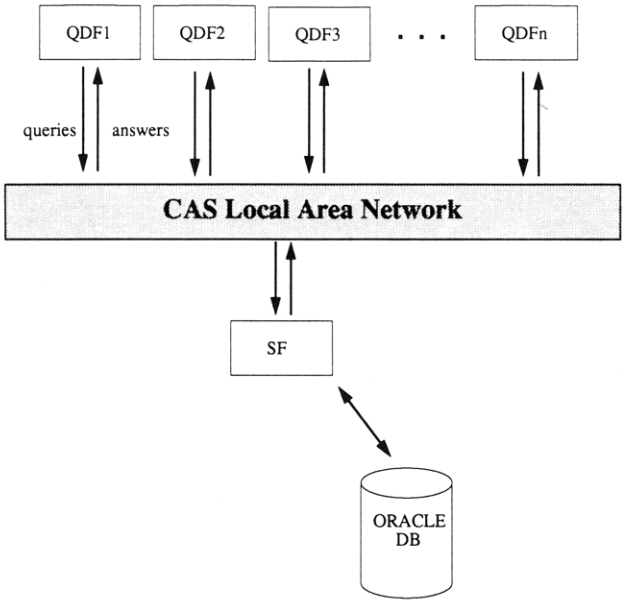


Figure 1. Initial computer configuration for SIM3D/ESS at CAS.

Programming Language). The functionality of a QDF relative to 3D substructure searching is described in a later section.

The search facility server is the search engine for 3D substructure searching. Currently, the search facility is on a dedicated SUN SPARCstation-2. The server accepts a query as a modified SYBYL MOL file or MOL2 file from a QDF client via a local area network (LAN). It executes various searches and retrieves, via the network, 3D database answer structures as SYBYL MOL(2) files to the QDF client for display and manipulation. A relational database management system (ORACLE) is used by the server for storage and retrieval of 3D structures.[39] The search facility can also be used to formulate simple queries and manipulate 3D structures, such as calculating the distance between a pair of atoms and displaying the connectivity and coordinates. An illustration of some the search facility commands is given in a later section.

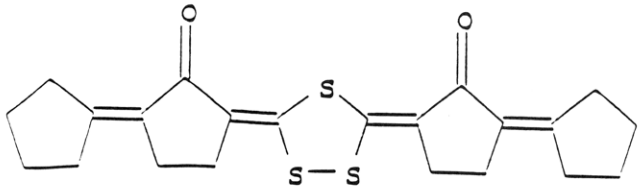
Figure 1 illustrates the current computer configuration for the SIM3D/ESS at CAS. Several QDF clients are shown connected to the CAS LAN. The arrows indicate the direction of the query and answer flow.

III. DATABASES

The initial databases for the SIM3D/ESS are current or potential subsets of CAS's new licensed file product, CAS 3D structure templates (CAST-3D).[40] There are currently four SIM3D/ESS databases.

CAST-3D RIGID. CAST-3D RIGID is a CAST-3D subset of approximately 365 000 rigid or semirigid substances, i.e., substances with limited conformational flexibility for at least some portion of their structure. There is a considerable amount of diversity among the substances. Substances on CAST-3D RIGID are useful as templates in syntheses since they can "lock-in" a desired geometric orientation of potential substituent atoms because of their rigidity.

The determination of flexibility for CAST-3D RIGID substances is partially based on a topological flexibility index called the global simple (GS) index that has been previously described.[18] This index is an average of the local simple indices (LS) that are computed for each pair of nodes in the structure. The shortest path between a node pair is used and an approximation of the conformational flexibility is calculated



Cyclopentanone, 2,2'-(1,2,4-trithiolane-3-5-diylidene) bis[5-cyclopentylidene-

61656-32-4
2.68 GS Index

Figure 2. Example of a CAST-3D RIGID acyclic–cyclic substance.

2-Propenamide, 2-methyl-

79-39-0
2.20 GS Index

P,P-Diamidotetrathiohypophosphoric acid

79828-56-1
2.11

Figure 3. Examples of CAST-3D RIGID acyclic substances.

by considering the bonding character and the amount of branching along the path. In a sample of 6000 Registry substances, the GS values ranged from 1.35 (most rigid) to 11.73 (most flexible).[18] Normalized versions of the local simple and the global simple indices, i.e., LSN and GSN, respectively, are also computed to allow for a comparison of flexibility independent of path length. The LSN values are useful in 3D searching where the path length is typically unknown.

On CAST-3D RIGID 72% of the substances contain "extended" cyclics, i.e., ring systems with a rigid acyclic node adjacent to a ring. Specifically, these substances must have at least one ring system with at least one exocyclic multiple bond (double, triple, or tautomeric), excluding bonds to a chalcogen atom. The GS index value for the cyclic–exocyclic portion must be 2.4 or less, and the overall GS index must be 6.0 or less. An example of an acyclic–cyclic CAST-3D RIGID substance is shown in Figure 2. Approximately 2% of CAST-3D RIGID substances are pure acyclic substances. The selection criterion for including these substances is a GS index of 2.2 or less. Two examples of substances meeting this criterion are given in Figure 3. The remaining substances (26%) on CAST-3D RIGID are pure cyclics. These substances are ring system substances registered in the CAS Chemical Registry.

General Substance File. The general substance file (GSF) is a systematic sample of 60 000 substances taken from an earlier version of the Registry 3D structure file (every 75th substance out of approximately 4.75 million substances). Molecular property data were also generated for this entire set of substances. The GSF database is useful for "mirroring" the full 3D structure file.

Biologically Related Substances. The biologically related substances (BRS) database contains approximately 125 000 Registry substances that appear in substance collections that are related to biology. The locator field on the STN Registry File was used for the determination of the substance collections. For example, a CHEMLIST locator for a substance indicates it is on the EPA TSCA Inventory or subject to regulations under the Toxic Substances Control Act or similar legislation. The substances on the BRS database also have stereochemical information incorporated in the 3D structure at the node level.

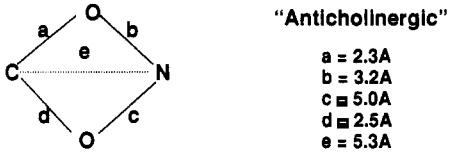


Figure 4. General 3D substructure search query: anticholineric pharmacophoric pattern.

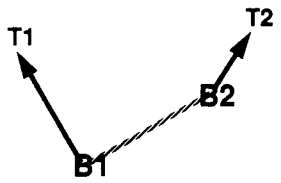


Figure 5. Bond vector pair: target of a 3D framework search.

This database should be especially useful for research on topics related to biology, such as drug design, toxicity studies, etc.

Chemical Source Subset. The chemical source subset (CSCHEMS) database contains approximately 30 000 substances. These are the substances having 3D coordinate information of the STN CSCHEM database. CSCHEM is a catalog file containing source information about chemical compounds and the firms that manufacture and/or distribute them.

IV. 3D SUBSTRUCTURE QUERY TYPES

SIM3D/ESS has been designed to support two logical types of 3D substructure search: general queries and framework queries. A general query focuses on searching a geometric pattern of atoms. Such queries are typically used to locate pharmacophoric patterns such as the "anticholinergic" pattern[41] illustrated in Figure 4 (a pharmacophore can be thought of as a geometric arrangement of atoms necessary to elicit a biological response). In specifying pharmacophoric pattern queries, qualifications such as global and local distance tolerances, H-donor and ring centroid node specifications, excluded volumes, etc., are often used. Most commercial 3D substructure search systems support this type of searching.

The amount of information that can be specified for a pharmacophoric pattern may, unfortunately, be minimal. It is significant, for example, that in the pattern shown in Figure 4 there are two atom triangles with interatomic distances. Sometimes only a single atom triangle is specified, but only rarely less information. The minimal unit in pharmacophoric pattern matching seems to be the triangle of atoms and their interatomic distances. This has been an important consideration in the design of SIM3D/ESS screen fragments.

The second type of 3D substructure query is a 3D *framework* query. In framework queries the focus is on the orientation of two or more bond vectors. The query in Figure 5 illustrates a bond vector pair target. The bonds are typically from a base atom (B) in a skeleton, such as a ring system, to a tip atom (T) of a substituent or a potential substituent such as hydrogen. This type of searching has been implemented in the CAVEAT system, developed by Bartlett et. al.[42] CAVEAT specifies the geometric relationship between a pair of bond vectors via the distance between the base atoms (B1, B2), the dihedral angle between the planes B1–B2–T2 and B2–B1–T1, and the angles T1–B1–B2 and T2–B2–B1. SIM3D/ESS exploits this relationship via special screens and atom-by-atom and superimposition matching discussed in the next section. Typically the objective in 3D framework searching is to locate a synthetic precursor template that will lead to a desired orientation of functional group substituents. Often it is desirable to "lock-in" the bond orientation in a rigid framework such as a ring system and, thus, minimize some of the design problems due to conformational flexibility.

It is important to note that the different substructure query types can be specified independently or together within the *same* query. However, special features are needed to appropriately handle each query type. For example, in framework searching, hydrogen atoms are legitimate vector tip targets. This requires special screens containing hydrogens and a more time-consuming atom-by-atom search where both the heavy atom skeleton and attached hydrogen atoms must be searched.

V. 3D SUBSTRUCTURE SEARCH CAPABILITIES

A. Query Framing. Query framing in the SIM3D/ESS is accomplished in the QDF using a special "ESS" menu for Tripos' SYBYL software. There are two logical modes of query framing: submodeling existing structures or building queries from scratch. In submodeling, a target 3D structure (or a portion of a structure) is used as the starting point for query creation. The query is usually specified as a node subset of the model, including its geometric relationship. Further modifications are made as appropriate. This mode is the most typical one for defining a 3D framework query.

In building queries from scratch, the nodes are input directly and appropriate constraints, such as distance and dihedral angles, are specified by the user. Currently, "native" SYBYL interaction is used to accomplish the query framing. This mode of query framing is typically used in specifying a 3D query for a pharmacophoric pattern.

The QDF currently can build queries with any combination of features: 2D paths and 3D points, vectors, bond angles, dihedral angles, excluded volumes, and included volumes. Specific, generic, and dummy nodes can be specified. Examples of specific modes are carbon and nitrogen; examples of generic nodes are halo and hetero; and examples of dummy nodes are lone pairs, ring centroids of planar five- and six-membered rings, and ring perpendiculars. The dummy atom nodes have been generated and added to the database 3D structures. Specification of global and local tolerances for distances, angles, etc., is also permitted. Also, a flexibility index range can be specified between a pair of nodes to help control the rigidity of the connecting path between them.

A 3D point is a three-dimensionally "significant" atom. Interatomic distances between 3D points can be automatically generated as query constraints. A 2D path defines the atoms of a bonded path whose atom and bond types are significant. Original bond types between the atoms are used during a search. However, the user may change these bond types to any specific or generic type. For example, 2D and 3D features may be combined to produce a constraint that is both three-dimensionally significant and a marker for desirable bond paths. Selected atoms retain their original atomic type when used in a query description. They may, however, be changed to a different specific atom type, to a generic atom type (e.g., halogen, hetero, or any), or to a variable group of atom types (e.g., G1 = O, N, P, S). In addition, the atom connectivity and fusion may be specified (a fused atom is one with three or more ring bonds).

A vector is a relationship between two 3D points. Distance constraints, flexibility constraints, and/or bond type constraints may be assigned to a vector. A distance constraint, composed of a minimum and a maximum distance, limits the allowable interatomic distance range for selected atom pairs. The flexibility constraint, also a range, limits the allowable

flexibility index values, via a tolerance, between two atoms in a molecule. It is primarily used to help control the conformational flexibility between a pair of nodes. The bond type constraint allows the user to specify the type of bond, if any, that should connect the two atoms in a vector.

An angle is used to relate three three-dimensionally significant atoms by the angle that two interatomic distances form. A minimum and maximum angle value must be assigned to each angle when it is created. Likewise a dihedral angle is used to relate four three-dimensionally significant atoms by the dihedral angle that they form. A minimum and maximum torsional angle must also be assigned to each dihedral when it is created.

Volumes (either included or excluded) define spherical regions in which answer structures must or cannot exist. The radius of each sphere is user-defined. Different volumes in a query may have different radii. All volumes are referenced by the atoms defining their centers. They are defined by (1) selecting an atom as the center of a volume and defining a radius, (2) selecting three (or more) atoms to define a plane and then specifying the perpendicular distance away from the plane and the volume's radius, or (3) selecting the *xyz* coordinates of the center and defining a radius.

B. Query Editing. Both the QDF and the search facility provide fairly extensive query editing. The QDF has menus which allow the user to change or reset atom types and vector, angle, dihedral, and volume constraints. Also, an EDIT command on the search facility server allows the user to change the atom types, bond types, interatomic distances and tolerances, and normalized flexibility index ranges in the query. However, unless the query is transmitted to the QDF client, the changes made with the EDIT command are not reflected during answer set display.

C. Answer Display. In viewing retrieved 3D structures on the QDF, the atoms in the file structure that matched the query atoms are "highlighted" with a contrasting color. Other display features are handled via menu commands. Typically, a simultaneous 3D display of the answer structure on top of the query is used in viewing an answer set, i.e., a superimposed solution. Different solutions may be viewed one at a time or in groups of up to four answers at a time. Menu options allow sequential or direct access viewing of the answer set by answer number or Registry Number. Two-dimensional views of answer structures may also be displayed.

Menu options also allow inspection of answer structures, including interatomic distances, angles, dihedral angles, and local and global flexibility values.

D. Search Levels. Screen search is typically the initial search level. The screens used in searching are primarily predetermined characteristics or features of 3D structures. They eliminate file substances not having the necessary corresponding query characteristics from further, more precise, but more time-consuming searching. Screen search in the SIM3D/ESS is implemented via inverted bitmap processing, where the screen bitmaps contain the database substance numbers having a particular screen. For the candidate substances passing the screen search level, an optional atom-by-atom search is performed on the 2D, 3D, or 2D/3D structure graph using a modified Ullmann subgraph isomorphism algorithm.[43] The Ullmann modifications include integrated 2D/3D searching, handling atom hybrids, Cheng and Huang performance heuristics,[44] and the reduction of memory usage by approximately half.[45] Users can specify whether the atom-by-atom search retrieves the first solution, all solutions, or the best solution per query/candidate answer

```
ESS> search

(1)  Tolerance for unspecified distances in angstroms (default=30.0):
(2)  Match on hybrid atom types? (default=no):
(3)  Match on query H-bonding attributes? (default=no):
(4)  Enter normalized global flexibility range (n - m)? (default=none):
(5)  Current answer set has been cleared!
(6)  Screen search in progress...
(7)  457 screened answers found
(8)  Maximum number of answers? (default=all):
(9)  Maximum structure size (heavy atoms)? (default=none):
(10) Exact, substructure, or superstructure search? (default=substructure):
(11) First, all, or best solutions? (default=first):
(12) Begin search? (default=y):
(13) 3D substructure atom-by-atom search in progress...
(14) Search 10% complete:  2 answers and 2 solutions found
(15) Search 25% complete:  3 answers and 3 solutions found
(16) Search 50% complete:  4 answers and 4 solutions found
(17) Search 75% complete:  6 answers and 6 solutions found
(18) Search 100% complete: 8 answers found
```

Figure 6. SEARCH command parameters for a sample search.

(as there may be more than one solution per file structure).

The final flexibility index range searching is accomplished in conjunction with atom-by-atom search. For file substances passing atom-by-atom search, the local and/or global normalized flexibility indices are generated and compared to the range specified in the query for the appropriate nodes and/or for the entire molecule. Volume, atom connectivity, fusion, and angle constraints in the query are also checked at this time.

File substances passing atom-by-atom search are superimposed onto the query using a root-mean-square-error (RMSE) superimposition fit provided that 3D coordinates are available for the query. The RMSE values are used to rank the answer set. Our superimposition algorithm is based on the method of Sippl and Stegbuchner.[46]

E. Search Facility Commands. The search facility provides a set of commands for uploading a query from the QDF client, searching the query against a selected database, tabular viewing of the query and matching file structures, ranking of the answers, and downloading an answer file to the QDF client for graphical display and manipulation. Simple queries may be constructed at the search facility using a nongraphical interface.

Atom-by-Atom searching supports several search options: substructure (inclusion-match) search, full structure (identity-match) search, and "superstructure" search can be specified. Superstructure search is an inverse substructure search; i.e., the file structure needs to be embedded in the query. This is implemented by reversing the roles of the query and file structures in the Ullman search routine. Subset searching using an answer set as an input parameter is also available.

A sample GSF database search is presented in Figure 6. All the default parameters were taken. Lines 1–7 apply to the screen search level. Line 1 gives the tolerance for any *unspecified* interatomic distance as 30 Å. These are interatomic distances automatically generated from 3D points. For example, in a query with three atoms and two distances, the third distance is assumed to have a tolerance of ±30 Å (i.e., this distance constraint is "logically" ignored during searching). Also, if a distance vector is specified without a tolerance, the tolerance given here is used, (i.e., line 1 can be used to specify a global tolerance). Line 2 specifies that the

Table I. Major Screen Classes

generic 2D fragments	103
atom pair distances	1430
atom triangle distances	9016
atom tetrangle distances	2846
flexibility indices	1124
total	14519

hybridization attribute of atoms such as "sp3" be ignored in the matching of query atoms. Line 3 specifies that query H-bonding attributes such as H-donors derived from atoms in the query not be used. However, explicit H-bond donors and acceptors (atom types HD or HA) are still used in searching. Line 4 specifies that there is no global flexibility index range constraint. If a range had been given, then this information would be used to generate the global flexibility screen(s) as a constraint for atom-by-atom searching. Line 7 reports on the number of candidates passing screen search. Currently, there are no screen search or atom-by-atom search limits. Lines 8–18 apply to the atom-by-atom search level. Line 8 indicates that there is to be no termination of searching after locating a certain number of answers. Line 9 specified that there is no heavy atom minimum threshold. Line 10 specifies the substructure search type, and line 11 specifies the file structure is a hit if just one solution is located. Lines 13–18 report on the progress of atom-by-atom searching.

F. Answer File Manipulation. The answer file manipulation capabilities in the search facility include the ranking of answers by RMSE value (superimposition), common-atom percentage (of query atoms to file structure atoms), or difference in the 3D Wiener index[47] (superimposition ranking is the default at search time). The interatomic distance matrix for the retrieved substances is used in calculating the 3D Wiener index. Answer files can be downloaded from the search facility, and Boolean operations among answer files can be applied, i.e., AND, OR, XOR, and NOT. Also, as mentioned above, answer files can be used as search parameters for subset or offspring searching.

Another important feature in the SIM3D/ESS with respect to answer files is the ability to "crossover" Registry Number answers to the STN REGISTRY, CA, and CASREACT Files to obtain additional information concerning structures in the answer set. The REGISTRY File can be searched to obtain chemical names, ring analyses data, 2D structure diagrams, and recent bibliographic reference data. 2D substructure searching of REGISTRY can be used to obtain related answer sets. The results of REGISTRY searches can be used to formulate queries for STN MARPAT, which can provide information about Markush structure claims in patents. The CA File can be searched for substance uses, properties, etc., and for a comprehensive set of bibliographic, index, and abstract data. CASREACT can be searched for synthetic information on substances. This is especially useful for the 3D framework searching results, where the location of synthetic precursors is often the objective of the search.

G. Search Screens. Effective search screens are crucial if files of several million 3D structures are to be searched efficiently. The current SIM3D/ESS screen set contains approximately 14 500 screens. There are five major classes of screens. Table I gives the number of screens in each class.

The purpose of the generic 2D fragment class is to improve the effectiveness of screen searching for queries with 2D character. The screens consist of element counts (such as a count of eleven or more carbons), degree of connectivity counts (such as a count of three or more of an atom with a connectivity of four, i.e., four non-hydrogen attachments), and bond counts (such as eight or more chain single bonds). There are a total

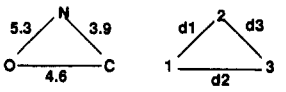


Figure 7. Derivation of atom triangle "5-slot" screens.

Table II. Atom Triangle Screen Classes

atom 1–atom 2	no. of screens	atom 1–atom 2	no. of screens
1. C–C	1728	7. N–Ht	125
2. O–C	1331	8. Hd–C	729
3. N–C	729	9. Hd–Ht	729
4. O–Ht	729	10. Ha–C	729
5. Oh–Ht	729	11. Ha–Ht	729
6. Oh–C	729		

of 103 screens in this class. The addition of more precise 2D fragment screens is planned for the near future (i.e., in the next version of the screen set). These will include primarily augmented atom fragment screens. (An augmented atom is a central atom and its nearest neighbor atoms).

Atom-pair distance screens are the traditional screens for 3D substructure searching. We use a total of 1430 screens in this class. Specific atom pairs are included such as C–N, N–O, and H–H [the H—H pair is useful in framework searching (*vide infra*)]. There are also generic atom pairs such as C–ANY, hetero–hetero, and N(SP3)–ANY. Hybridization attributes are available for only one atom of the pair; the other atom must be ANY. Dummy atoms such as lone pairs and ring centroids are also included in the atom-pair screens.

The interatomic distances for the atom pairs are "binned" into 14 or 26 bins, depending on their frequency of occurrence. More frequently occurring atom pairs such as C–N receive 26 bins, while less frequently occurring pairs such as I–Br receive 14 bins. The Lederle binning scheme is used for the 14 bins.[9] We have further subdivided these bins to obtain a 26-bin set. Some examples in the 26-bin set are bin-2 = 0.480–0.898 Å, bin-6 = 1.910–2.147 Å and bin-24 = 11.008–21.109 Å.

The third major class of screens is a novel set of atom triangle distances. We have defined 9016 screens for this class. The triangle screen fragments are represented as "5-slot" keys. Two of the three atoms in the triangle and the three distances are included in the key. Figure 7 illustrates the derivation of 5-slot keys.

The atoms in the triangle are ordered by atomic number. In this example triangle, oxygen is the atom most preferred (atomic number 8) followed by nitrogen (atomic number 7) with carbon (atomic number 6) as the least preferred atom. The D1 distance is the distance between atoms 1 and 2, D2 the distance between atoms 1 and 3, and D3 the distance between atoms 2 and 3. The triangle screen fragment is the first atom followed by the second atom followed by D1, D2, and D3 in sequence. Atom types in 5-slot keys may be a specific atom such as N or a generic atom such as Ht (a hetero), Oh (a hetero other than O or N), Hd (an H-donor), and Ha (an H-accepter). Table II gives the At1_At2 atom classes and the number of screens in each class.

The three interatomic distances are placed into an equal number of bins, depending on the relative frequency occurrence of the atoms in the triangle. Five, nine, eleven, and twelve equifrequency bins were used for D1, D2, and D3 (per atom pair). For example, the number of screens for the C–C preferred atom pair in a triangle is 12^3, i.e., all possible combinations of the 12-bin D1, 12-bin D2, and 12-bin D3

C C (At1-At2)

	D1	D2	D3
1.	0.0 - 1.8	0.0 - 1.8	0.0 - 2.2
2.	>1.8 - 2.1	>1.8 - 2.1	>2.2 - 2.9
3.	>2.1 - 2.5	>2.1 - 2.5	>2.9 - 3.7
4.	>2.5 -3.2	>2.5 - 2.9	>3.7 - 4.2
5.	>3.2 - 3.8	>2.9 - 3.4	>4.2 - 4.7
6.	>3.8 - 4.4	>3.4 - 3.9	>4.7 - 5.2
7.	>4.4 - 5.0	>3.9 - 4.4	>5.2 - 5.8
8.	>5.0 - 5.8	>4.4 - 5.1	>5.8 - 6.4
9.	>5.8 - 6.6	>5.1 - 5.8	>6.4 - 7.0
10.	>6.6 - 7.7	>5.8 - 6.4	>7.0 - 7.8
11.	>7.7 - 8.7	>6.4 - 7.5	>7.8 - 8.6
12.	> 8.7	> 7.5	>8.6

Figure 8. Illustration of distance bins for C–C preferred atom pair of an atom triangle.

sets. The nature of the distance ranges of the bins for the C–C preferred atom pair in a triangle is illustrated in Figure 8.

Note that the distance ranges for a particular bin can vary between D1, D2, and/or D3. For example, for bin-6, D1 is 3.8–4.4 Å, D2 is 3.4–3.9 Å, and D3 is 4.7–5.2 Å. All combinations may be used, but in practice some combinations are not used or occur very infrequently. For example, for the GSF database, only 8877 screens out of the 14 519 possible are used. This is due to the use of "calculated" atom triangle and tetrangle screens (*vide infra*). The much larger CAST-3D RIGID database uses only 9703 screens; i.e., it does not have a proportionally larger number. The BRS database uses 8831; the CSCHEM database uses 8679. We plan to streamline the screen set in the next version by "overlapping" (logical OR) into a single screen several rarely occurring triangle and tetrangle fragments.

The purpose of the triangle fragment keys is to capture as much of the information inherent in the triangle as possible. Triangles, as mentioned earlier, constitute an important unit in pharmacophore pattern matching. In our preliminary testing involving a sample of 10 "benchmark" triangle queries, the triangle-based screens have significantly improved screening relative to using only atom pair distances. These triangle queries are a systematic sample of some 15 million triangles that were generated for a sample of 6K substances. They are illustrated in Figure 9. The queries are fairly generic relative to typical pharmacophoric pattern queries; i.e., they tend to be "worst case" queries. For example, 6 out of the 10 queries, two of which are all carbon (nos. 6 and 8) have one or no hetero atoms.

Figure 10 illustrates the screen efficiency for the sample triangle queries using the appropriate combination of atom pair *and* atom triangle screens in our screen dictionary. Percent screen efficiency is defined as the ratio of atom-by-atom search hits to the screen search hits times 100. For example, if the numbers of atom-by-atom search hits and screen search hits are the same, then the efficiency of the screen set is 100%; if the number of screen hits is twice the number of atom-by-atom hits, the efficiency is 50%, etc. The searches were run at four different global distance tolerances: 0.05, 0.20, 0.50, and 1.0 Å. The "square", "plus", "star", and "diamond" symbols correspond to tolerances of 0.05, 0.20,

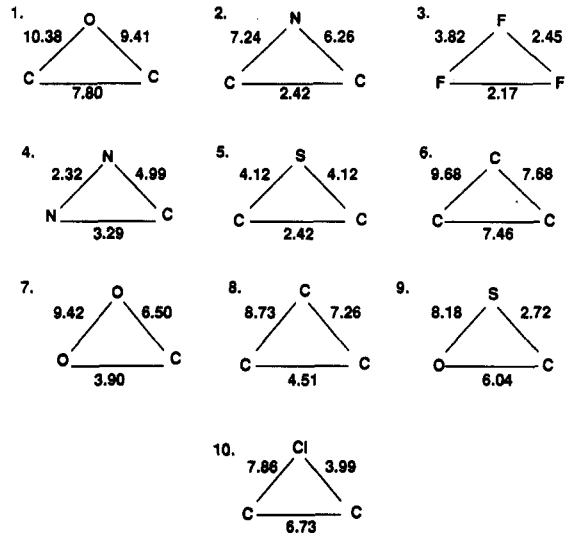


Figure 9. Sample atom triangle queries.

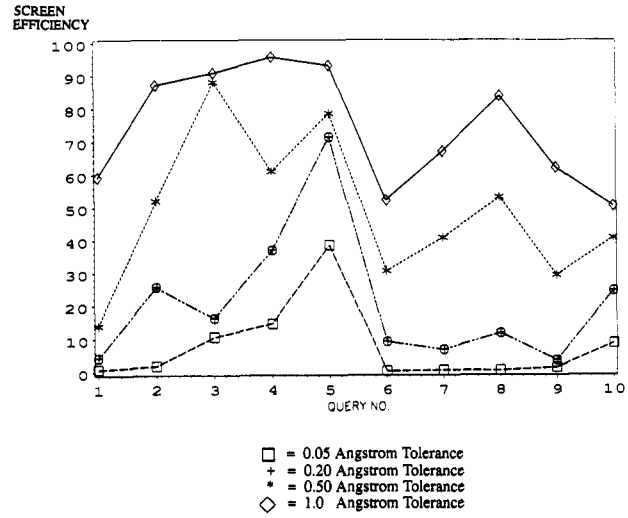


Figure 10. Screen efficiency of atom pair and atom triangle screens for 10 sample triangle queries using 0.05-, 0.20-, 0.50-, and 1.0-Å global tolerances.

0.50, and 1.0 Å, respectively. It is important to note that the tolerances are for each side of the triangle. Figure 10 shows that the screens are more efficient when the global tolerance is the largest. This has been one of our design goals, since larger tolerances are probably more typical in queries than the smaller tolerances. The greater screen efficiency at the larger tolerances is probably due to the greater correspondence of tolerance to the typical bin width. For example, the mean bin width for the triangle atom pair illustrated in Figure 8 is 0.69 Å for D1, 0.67 Å for D2, and 0.64 Å for D3, excluding the first and last bins. To improve the efficiency at lower tolerances a "finer" binning scheme would be needed, i.e., with narrower bin widths. This would result, of course, in a significant increase in the number of possible screens.

There is, however, an opposite trend with respect to percent screenout (which is defined as the ratio of the difference between the number of structures in the file and the atom-by-atom search hits, and the number of structures in the file times 100). The greater the percent screenout, the less time spent atom-by-atom searching. Figure 11 gives the screenout for the sample triangle queries corresponding to the screen efficiency given in Figure 10. The decrease in percent screenout with increase in global tolerance is observed, but the amount of change for the queries is not as significant as that for the screen efficiency.

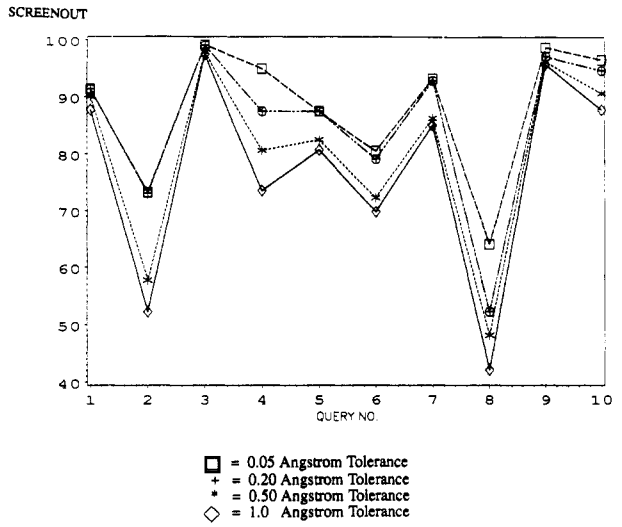


Figure 11. Screanout of atom pair and atom triangle screens for 10 sample triangle queries using 0.05-, 0.20-, and 1.0-Å global tolerances.

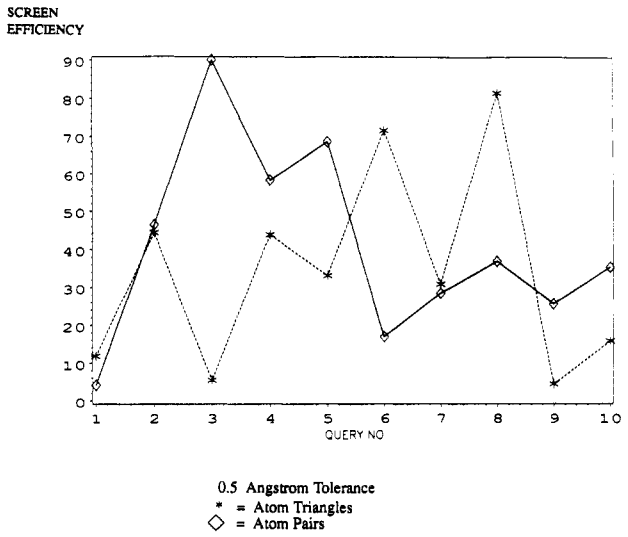


Figure 12. Screen efficiency of atom pair vs atom triangle screens for 10 sample triangle queries at 0.50-Å global tolerance.

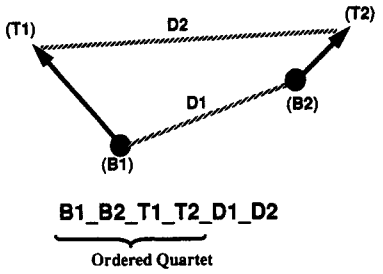


Figure 13. Derivation of atom tetrangle screens.

The complementary interaction of the atom pair and atom triangle screens is illustrated in Figure 12 as a plot of the percent screen efficiency of atom pair screens *versus* the atom triangle screens for the sample triangle queries. The two screen classes tend to compensate for each other; i.e., if efficiency is low for one type, it tends to be high for the other type and vice versa.

The atom pair and atom triangle screens are targeted primarily for general 3D substructure searching. We have designed a third major class of screens which we call bond vector "tetrangles". These tetrangles are targeted primarily for 3D framework searching. Figure 13 illustrates the derivation of a tetrangle screen fragment. The base atoms, B1 and B2 in the illustration, can be a carbon, a hetero, or an any; the tip atoms, T1 and T2, can be a carbon, a hetero, a hydrogen, or an any. D1 is the distance between the base atoms; D2 is the distance between the tip atoms. As with triangle fragments, ordered tetrangle fragment keys are used. The base atoms, B1 and B2, are ordered by atomic number; the tip atoms, T1 and T2, are ordered but are independent of the base atoms (i.e., the correspondence between the base atom with and tip atom is not maintained in the key).

Table III. Atom Tetrangle Screen Classes

	B1	B2	T1	T3	freq, %	no. of screens
	ANY	ANY	ANY	ANY	100	400
	C	C	H	H	7.4	1296
	C	C	C	H	24.5	225
	C	C	Ht	H	5.1	400
OR	C	C	Ht	Ht	1.1	
	C	C	C	C	29.2	225
	C	C	Ht	C	9.6	100
	Ht	C	C	H	6.4	100
OR	Ht	C	Ht	H	1.8	
	Ht	C	H	H	0.6	100
OR	Ht	C	Ht	C	5.3	
	Ht	C	Ht	Ht	0.8	
	Ht	Ht	H	H	0.03	
	Ht	Ht	C	H	0.3	
	Ht	Ht	Ht	H	0.09	
	Ht	Ht	Ht	C	0.7	
	Ht	Ht	Ht	Ht	0.2	

D1 and D2 are equifrequency binned with either 10, 15, 20, or 36 bins, depending to some extent on the frequency of occurrence of base–tip atom combination. The tetrangle screen classes, their frequency of occurrence (on a random sample of 500 substances), and the corresponding screens per class are shown in Table III. Three of the classes illustrated in Table III are shared combinations of several more specific classes connected by Boolean OR logic.

In general, the number of bins per tetrangle fragment or fragment group was based on the fragment's frequency of occurrence. For example, when both the base and tip atoms are any atoms, 20 bins each per D1 and D2 are used. Thus, there are 20^2 or 400 screens when all four of the atoms are any. However, the number of bins for (H)C–C(H) and (Ht)C–C(H) or (Ht)C–C(Ht) are significantly greater than for other tetrangles. The reason for the use of 1296 screens for (H)C–C(H) is that the fragment is prevalent in queries and also because there are no supporting triangle screens with H atoms. The reason for the other case is anticipated prevalence in queries.

The tetrangle screens are not as precise as those used by Bartlett,[42] who used four parameters to describe a vector pair: the base atom to base atom distance, the two angles of the vectors with the base, and the four atom dihedral angle. However, our tetrangle screens are supported by atom pairs and atom triangle screens, and our screen search is followed by the precision-refining atom-by-atom and superimposition searches, whereas CAVEAT has only a single search level. Thus, our final answer sets should be as precise or more precise than those obtained using CAVEAT.

Flexibility index screens are the fourth major class of screens. There are a total of 1124 screens in this class. Fifty of these screens are allocated on an approximately equifrequency basis for the global simple normalized index of a substance. This index gives an approximate measure of the overall conformational flexibility of the substance. An example of a GSN screen is 0.361–0.356, which accounts for 2% of a set of sample substances. Also, 50 screens are similarly allocated for the global simple index of a substance. The majority of the

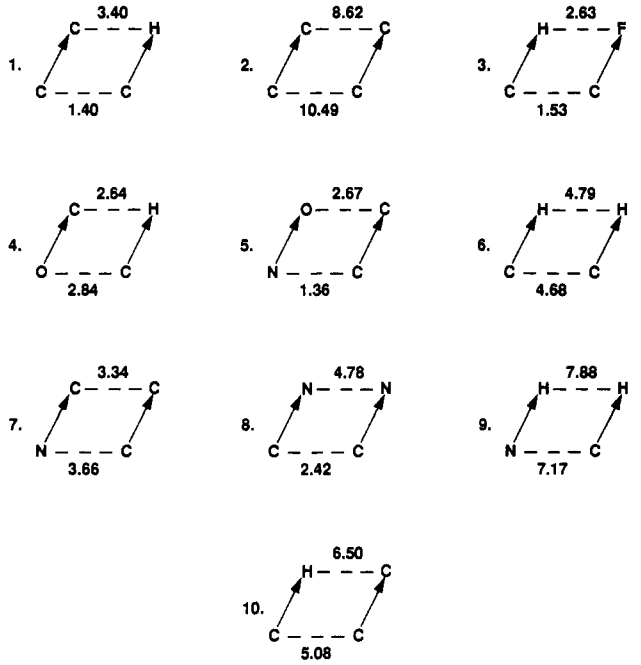


Figure 14. Sample tetrangle queries.

Table IV. Results of Tetrangle Searches with Screen Set

Qno	screen hits	screenout	% screen efficiency	screen hits (local flex)	screen hits (global flex)
1	5563	7.0	98.6	–	–
2	2074	65.3	19.5	363	63
415	3	93.1	63.9	–	–
4	3073	48.6	58.7	–	–
5	643	89.2	87.7	–	–
6	3570	50.5	26.5	676	82
7	2344	60.8	49.1	270	40
8	598	90.0	65.7	–	–
9	1059	82.3	5.8	76	13
10	3835	35.9	41.2	756	118
mean		62.3	51.8		

flexibility screens (1024), however, are allocated for the local simple normalized (LSN) indices of a substance. Unlike the GSN screens, where one screen is set per substance, multiple LSN screens are set per substance, i.e., one per atom pair, albeit more than one atom pair may map to the same screen. The LSN screens are subdivided into four classes corresponding to a subset of those used for the atom pair distance screens. These classes include the cases where the terminal atoms are (1) both specific hetero atoms such as N—O, (2) a carbon to specific hetero atom such as C–O, (3) both carbons, and (4) both generic "ANY" atoms. An example of a LSN screen is 0.062–0.075, which accounts for 2.7% of a set of sample substances. Final resolution of a global or local flexibility specification is accomplished at the atom-by-atom search level.

To test the effectiveness and efficiency of the screen set for framework queries we used a systematically selected set of "tetrangle" queries from the tetrangles generated for a set of 500 randomly selected GSF substances. These queries (see Figure 14) are fairly generic with several D1 distances within bonding range and/or with H tip atoms. The "arrows" between the atoms indicate the presence of "any" bond. The results of searching for these queries against a test file of approximately 6K GSF substances and using 0.2-Å tolerances are given in Table IV.

The screen set provides excellent efficiency as indicated by the mean-percent screen efficiency of 51.8%. However, the overall screenout is low due to the generic nature of the queries. Despite the high screen efficiency, some of these searches would require a considerable amount of the time-consuming atom-by-atom searching, unless they are augmented with additional query features or search parameters. The query with the lowest screen efficiency is no. 9, which has two H tip atoms, and consequently, does not use atom triangle screens—H atom triangles are not generated. For the next version of the screen set we plan to use a larger number of screens for the Ht–C/H–H fragment (see Table III). This should significantly improve the screen efficiency for queries with this fragment.

One way to modify queries to improve screenout is to use flexibility screens. The next to the right-most column in Table IV gives the number of screen hits when a local flexibility between the base atoms was specified at 0.6–1.0; i.e., the base atoms must be in a rigid environment. The specification was made only for queries with a distance greater than 3 Å between the base atoms. Table IV shows a significant reduction in the numbers of screen hits for these queries. When a global index range of 0.6–1 was specified for these queries, searching was restricted to file substances whose overall structure is rigid. The specification of global flexibility reduces the screen hits by even a greater amount than the local specification (see last column in Table IV).

VI. APPLICATION SCENARIO

This section describes an application scenario involving a 3D framework query. The application goal is to design a carbonic anhydrase inhibitor for use as a topical antiglaucoma agent. Reduction of intraocular pressure (IOP) is the primary method of treating blindness caused by glaucoma and can be achieved by complete inhibition of the enzyme carbonic anhydrase.[48] Orally administered carbonic anhydrase inhibitors (CAI), such as azetazolamide, have some effect but also have systemic side effects. Topically administered CAI's are needed to avoid some of these side effects and deliver a dose directly to the eye. Unfortunately, a reasonable formulation of a CAI is difficult to obtain since most solid sulfonamides (the most common CAI) are sparingly soluble. One must also take into consideration penetration of the corneal membrane in the design of a new agent (i.e., they must be reasonably lipophilic).

A few years ago a Merck research team developed several potential topically active CAI's, including an isobutylamino theinothiopyransulfonamide called MK417 (CAS Registry Number: 123308-22-5).[48,49] Our molecular design objective was to postulate substances with characteristics, especially geometric ones, similar to Merck's MK417. Our procedure was to use the coordinates of the carbonic anhydrase bound with azetazolamide[50,51] and to dock the prototypical MK417 into the active site according to the positioning described in ref 48. Figure 15 illustrates the overlap of the MK417 model onto the bound configuration of the azetazolamide inhibitor using SYBYL.[38] The isobutylamino side chain on MK417 is in the upper right-hand portion of carbonic anhydrase cavity.

The derived 3D framework query is shown in Figure 16. Vector pairs were used to orient the substitution point for the sulfonamide "warhead" and a N-isobutyl group that surprisingly increased the potency of MK417. A ring centroid for a planar ring was included to ensure that the sulfonamide would be attached to an aromatic ring in any resulting answers. The tip atoms were converted to H's and the base atoms to ANY's in the query. Tolerances of 0.1 Å were specified about each interatomic distance. A maximum of 20 heavy atoms was specified as well.

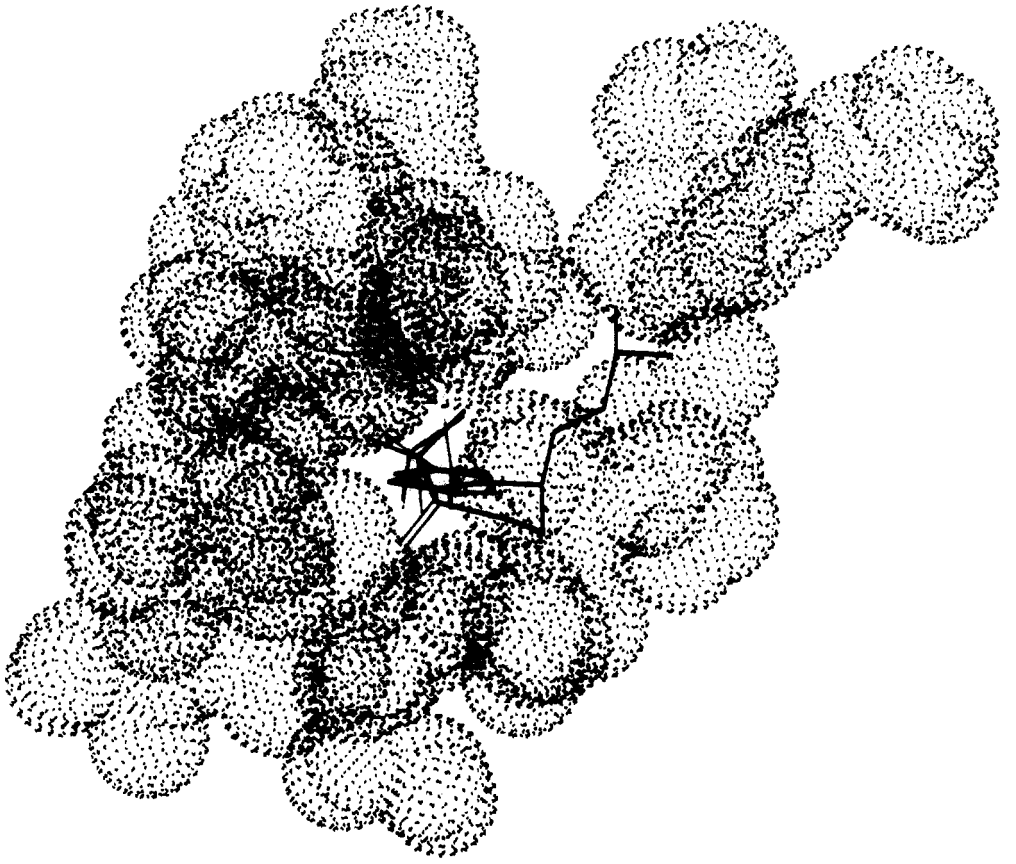


Figure 15. Overlap of the prototype carbonic anhydrase inhibitor, MK417, onto the bound configuration of a known inhibitor, acetazolamide.

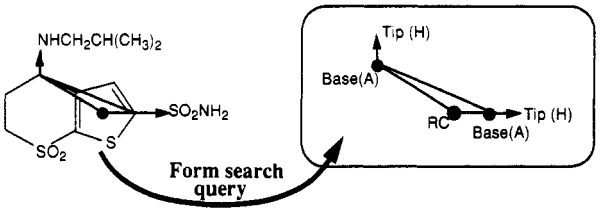


Figure 16. 3D framework substructure search query based on MK417.

A 3D substructure search using the query was executed against a file of 60K ring system substances, which was the precursor file to CAST-3D RIGID. [It is important to note that some of the ring system substances on this database have been registered for system purposes and serve as "ring parents" for substances containing closely related ring systems which are cited in the literature.] The purpose of this ring system database search was to locate rigid templates with correctly oriented substituent possibilities. The hit templates were then docked in the active site and appropriate substituents were modeled.

A total of 97 template answers were retrieved. Some examples of these are shown in Figure 17. We have further explored some substance design possibilities based on CAS Registry Numbers 327-06-0 and 115187-06-9. These are shown in Figure 18.

The similarity between the postulated CAI based on CAS Registry Number 327-06-0 and the prototypical MK417 is quite clear. However, the postulated substance based on the spiro template (RN: 115187-06-9) was the result of a serendipitous discovery. The actual match of one of the vector tips was to a hydrogen next to the spiro ring. Additional examination of this template indicated that the spiro ring could "lock-in" the orientation of its isopropyl group in the desired position relative to the isobutylamino group on MK417.

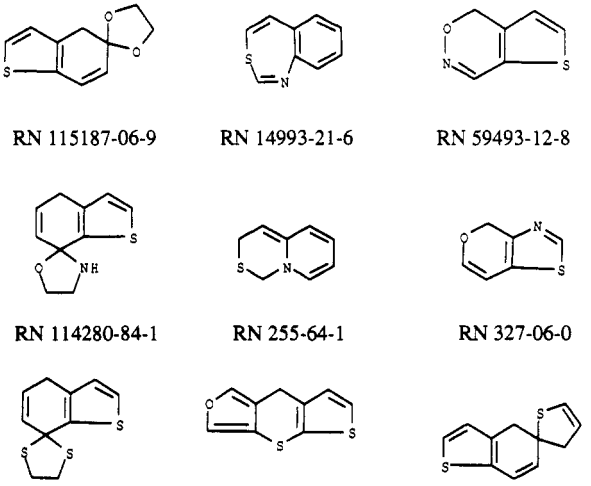


Figure 17. Sample retrievals from 3D framework search on ring system database.

Figure 19 shows that the CONCORD-generated geometry of the designed spiro substance fits well in the active site pocket. The designed substance (bold lines) is overlapped on the MK417 prototype. Further optimization of the geometry of the designed substance in the active site revealed no unexpected structural changes.

To obtain an estimate of the transportability of the possible designed substances based on the ring templates shown in Figure 17, the log *P* values for each template and other important fragments were estimated by using the atomic hydrophobicity method of Ghose et. al.[52] The estimated log *P* values are given in Table V. As seen in Table V, there are templates in the range of a phenyl group log *P*. With appropriate substitutions, these postulated CAI's would be

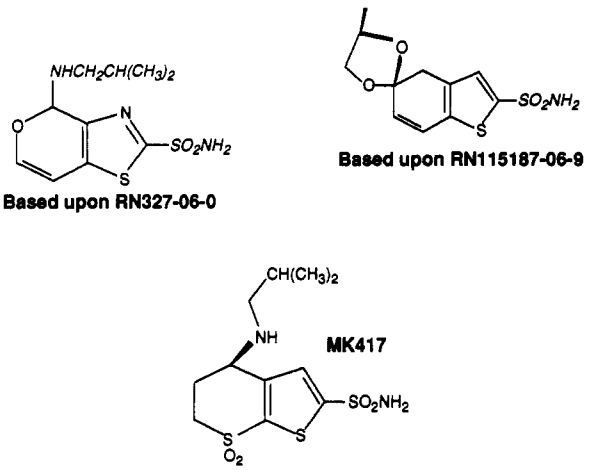


Figure 18. Possible carbonic anhydrase inhibitors based on MK417.

predicted to be as readily transportable across the corneal membrane as the prototypical MK417, i.e., based on the estimated log *P*. Since many sulfonamides are solids at room temperature, automatic estimation of their aqueous solubility becomes a nontrivial problem and was not attempted.

Synthetic pathways to certain candidates (either precursors or closely related compounds) are available through searching CASREACT and the CA File. Preliminary investigations using the structure search capabilities of STN indicated that neither of the postulated CAI's or related substances based on different degrees of saturation in the ring systems are currently found in the Registry File, indicating that the specific substances probably have not yet been patented. Subsequent search of the STN MARPAT file indicated that the substances were not part of a Markush structure claim.

Table V. Calculated log *P* Values for Ring Templates and Other Fragments

fragment or template	estd log *P* [52]
$-SO_2NH_2$	−0.67
$-NHCH_2CH(CH_3)_2$	1.04
disubstituted 1,3,4-thiadiazole	0.99
trisubstituted 1,3,4-thiadiazol-2-ine	0.74
disubstituted phenyl	1.28
thienothiopyran	0.05
RN327-06-0	−0.84
RN115187-06-9	0.69
RN114280-84-1	0.82
RN102601-68-3	1.05
RN255-64-1	1.22
RN66750-96-7	2.06
RN14993-21-6	2.62
RN120145-76-8	0.65
RN59493-72-8	−0.23

In conclusion, the above mentioned designed substances based on prototypical MK417 would seem to warrant further consideration in a drug design process for topically active carbonic anhydrase inhibitors.

VII. SUMMARY AND CONCLUSIONS

The CAS experimental system is unique in its handling of both general and framework 2D/3D substructure search queries in a fully integrated manner. The general query type is typically used for pharmacophore pattern matching, while the framework type is typically used in locating synthetic precursors leading to a desired geometric orientation of substituents. Both a "build" and a "submodel" mode of query framing are supported. In query framing, 2D paths, 3D points, vectors, angles, dihedral angles, and included and excluded

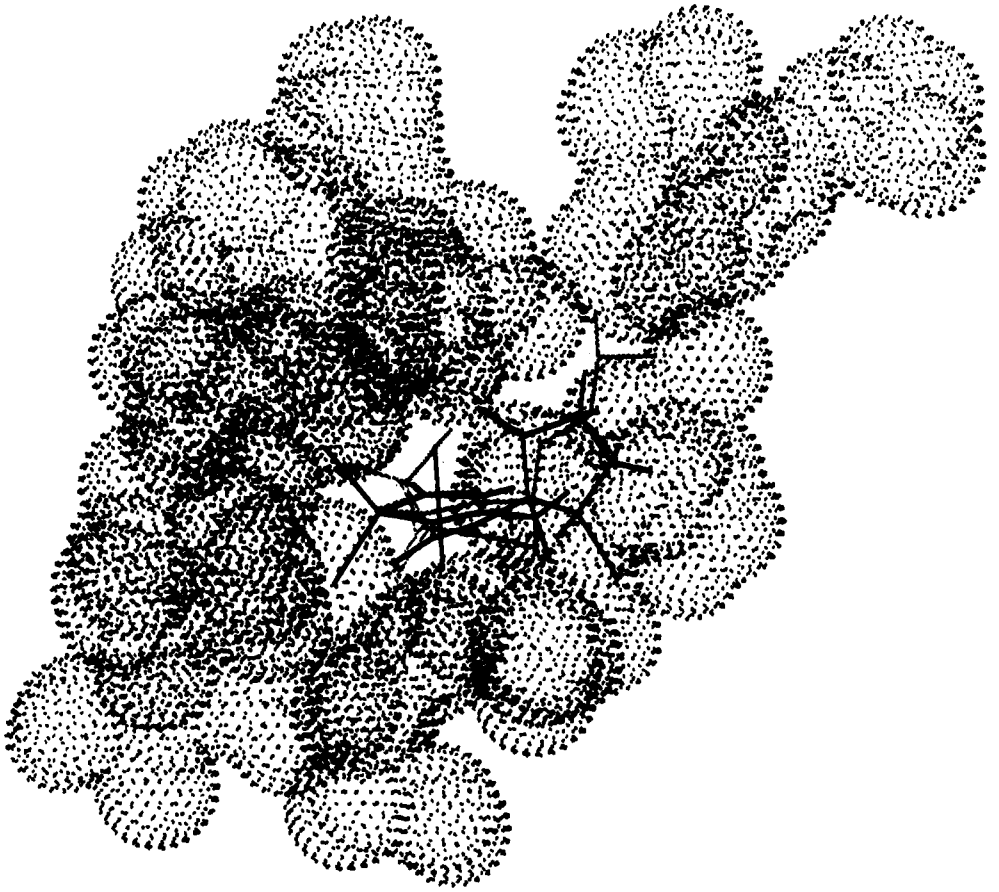


Figure 19. Overlap of MK417 and designed spiro substance in active site of carbonic anhydrase.

volumes may be specified. Specific, generic, dummy, (such as a lone pair or a ring centroid), and variable atom types may be used.

Four databases available for searching provide access to a fairly diverse set of over 560 000 substances. CAST-3D RIGID is a useful target especially for framework queries, where the objective often is to locate templates which can "lock-in" a desired orientation. The general substance file tends to mirror the full 3D Registry File. The biologically related substance file provides a set of substances incorporating stereochemistry information that may be useful in drug design, toxicology studies, etc. The chemical sources database is useful in locating purchasable synthetic precursors.

Searching can be executed at three levels: screen search, atom-by-atom search, including an integrated 2D/3D search, and superimposition fit, which checks the orientation of the matching atoms. Specification and searching global flexibility for a complete substance or local flexibility between atoms are novel and important features of the system. These allow users to control (to some extent) the overall conformational flexibility in the file substance and/or in the backbone connecting target atoms.

In addition to the traditional atom pair distance screens, we have also designed and use three novel types of 3D substructure screens: atom triangle distance, atom tetrangle distance, and flexibility index screens. On the basis of preliminary testing results, these four classes of screens are complementary in supporting 3D general and framework screening and seem to provide effective and efficient screening of fairly large files (500 000) of 3D structures.

We have also demonstrated how to develop a framework query from a 3D model of a known carbonic anhydrase inhibitor and to search for alternative templates in a file of rigid structures such as CAST-3D RIGID. Other CAS databases such as CASREACT and the CA File may then be examined to access synthesis information on the template itself and on the proposed substituted templates. The Registry, CA, and MARPAT Files may be used to obtain patent information on postulated substances.

We anticipate that adjustments to our 3D substructure screen set and search capabilities will be needed, reflecting new types of user queries and increasing performance for searching files of several million substances. Input to these adjustments is expected to result from an experimental service based on this system, which is to begin in 1993.

We plan to evolve the experimental system in the future by incorporating a conformational expansion search capability, where file substances passing a distance ranged screen and atom-by-atom search are analyzed to see if they can reasonably attain the geometry of the query. Other likely enhancements are the addition of molecular property databases and various fuzzy-match similarity search capabilities on the three data types, including speciality capabilities, such as shape/size and electrostatic similarity searching.

ACKNOWLEDGMENT

We would like to thank the following members of the CAS Research Staff for their assistance in this project: T. E. Bangert, D. P. Gieschen, A. H. Lipkus and G. G. Vander Stouw.

REFERENCES AND NOTES

(1) Gund, P.; Wipke, W. T.; Longridge, R. Computer Searching of a Molecular Structure File for Pharmacophore Patterns. *Proceedings of the International Conference on Computers in Chemical Research and Education*, Ljubljana; Elsevier: Amsterdam, 1974; Vol. 3, pp 5133–8.

(2) Gund, P. Three-Dimensional Pharmacophoric Pattern Searching. In *Progress in Molecular and Subcellular Biology*; Hahn, F. E., Ed.; Springer: Berlin, 1977; Vol. 5.

(3) Esaki, T. Quantitative Drug Design. V. Approach to Lead Generation by Pharmacophore Pattern Searching. *Chem. Pharm. Bull.* **1982**, *30* (10), 3657–61.

(4) 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* (1), 12–20.

(5) Jakes, S. E.; Watts, N.; Willett, P.; Bawden, D.; Fischer, J. D. Pharmacophoric Pattern Matching in Files of 3-D Chemical Structures: Evaluation of Search Performance. *J. Mol. Graphics* **1987**, *5* (1), 41–8.

(6) Brint, A. T.; Willett, P. Pharmacophoric Pattern Matching in Files of 3-D Chemical Structures: Comparison of Geometric Searching Algorithms. *J. Mol. Graphics* **1987**, *5* (1), 49–56.

(7) Martin, Y. C.; Danaher, E. B.; May, C. S.; Weininger, D. MENTHOR, a Database 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.

(8) 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 Substructure Searching of Three-Dimensional Molecular Structures. *J. Comput.-Aided Mol. Des.* **1989**, *3*, 225–51.

(9) 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*, 255–60.

(10) Moock, T. E.; Christie, B.; Henry, D. MACCS-3D: a New Database System for Three-Dimensional Molecular Models. In *Chemical Information Systems, Beyond the Structure Diagram*; Bawden, D., Mitchell, E. M., Eds.; Ellis Horwood: London, U.K., 1990.

(11) Cringean, J. K.; Pepperrell, C. A.; Poirrette, A. R.; Willett, P. Selection of Screens for Three-Dimensional Substructure Searching. *Tetrahedron Comput. Methodol.* **1990**, *1*, 37–46.

(12) Murrall, N. W.; Davies, E. K. Conformational Freedom in 3-D Databases. Techniques. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 312–6.

(13) Clark, D. E.; Willett, P.; Kenny, P. W. Pharmacophoric Pattern Matching in Files of Three-Dimensional Chemical Structures: Use of Smoothed Bounded Distances for Incompletely Specified Query Patterns. *J. Mol. Graphics* **1991**, *9*, 157–60.

(14) Poirrette, A. R.; Willett, P.; Allen, F. H. Pharmacophoric Pattern Matching in Files of Three-Dimensional Chemical Structures: Characterization and Use of Generalized Valence Angle Screens. *J. Mol. Graphics* **1991**, *9*, 203–17.

(15) Guner, O. F.; Henry, D. R.; Pearlman, R. S. Use of Flexible Queries for Searching Conformationally Flexible Molecules in Databases of Three-Dimensional Structures. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 101–9.

(16) Guner, O. F.; Henry, D. R.; Moock, T. E.; Pearlman, R. S. Flexible Queries in 3D Searching. 2. Techniques in 3D Query Formulation. *Tetrahedron Comput. Methodol.* **1992**, *3*, 557–63.

(17) Haraki, K. S.; Sheridan, R. P.; Venkataraghavan, R.; Dunn, D. A.; McCulloch, R. Looking for Pharmacophores in 3-D Databases: Does Conformational Searching Improve the Yield of Actives? *Tetrahedron Comput. Methodol.* **1992**, *3*, 565–573.

(18) Fisanick, W.; Cross, K. P.; Rusinko, A., III. Characteristics of Computer-Generated 3D and Related Molecular Property Data for CAS Registry Substances. *Tetrahedron Comput. Methodol.* **1992**, *3*, 635–52.

(19) Christe, B. D.; Henry, D. R.; Wipke, W. T.; Moock, T. E. Database Structure and Searching in MACCS-3D. *Tetrahedron Comput. Methodol.* **1992**, *3*, 653–64.

(20) Davies, K.; Upton, R. Experiences Building and Searching Chapman & Hall Dictionary of Drugs. *Tetrahedron Comput. Methodol.* **1992**, *3*, 665–71.

(21) Bures, M. G.; Hutchins, C. W.; Maus, M.; Kohlbrenner, W.; Kadam, S.; Erickson, J. W. Using Three-Dimensional Substructure Searching to Identify Novel, Non-Peptidic Inhibitors of HIV-1 Protease. *Tetrahedron Comput. Methodol.* **1992**, *3*, 673–80.

(22) Moon, J. B.; Howe, W. J. 3D Database Searching and *de novo* Construction Methods in Molecular Design. *Tetrahedron Comput. Methodol.* **1992**, *3*, 697–711.

(23) Rusinko, A., III; Skell, J. M.; Balducci, R.; Pearlman, R. S. *CONCORD User's Manual*; Tripos Associates: St. Louis, MO. Pearlman, R. S. Rapid Generation of High Quality Approximate 3-D Molecular Structures. *Chem. Des. Auto. News* **1987**, *2* (1), 5–6.

(24) Rusinko, A., III. Tools for Computer-Assisted Drug Design, Chapter 3. Ph.D. Thesis, The University of Texas, Austin, TX, 1988.

(25) Wigington, R. L. Machine Methods for Accessing Chemical Abstracts Service Information. *Proceedings of IBM Symposium on Computers and Chemistry*; IBM Data Processing Division: White Plains, NY, 1969.

(26) Fisanick, W.; Mitchell, L. D.; Scott, J. A.; Vander Stouw, G. G. Substructure Searching of Computer-Readable Chemical Abstracts Service Ninth Collective Index Nomenclature Files. *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 73–84.

(27) Dunn, R. G.; Fisanick, W.; Zamora, A. A Chemical Substructure Search System Based on Chemical Abstracts Index Nomenclature. *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 212–8.

(28) Farmer, N. A.; O'Hara, M. P. CAS ONLINE—A New Source of Substance Information from Chemical Abstracts Service. *Database* **1980**, *3*, 10–25.
(29) Zeidner, C. R.; Amoss, J. O.; Haines, R. C. The CAS ONLINE Architecture for Substructure Searching. *Proceedings of the 3rd National Online Meeting*; Learned Information, Inc.: Medford, NJ, 1982; pp 575–86.
(30) Dittmar, P. G.; Farmer, N. A.; Fisanick, W.; Haines, R. C.; Mockus, J. The CAS ONLINE Search System. 1. General System Design and Selection, Generation, and Use of Search Screens. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 93–102.
(31) Fisanick, W. Requirements for a System for Storage and Search of Markush Structures. In *Computer Handling of Generic Chemical Structures*; Barnard, M., Ed.; Gower: Aldershot, U.K., 1984; pp 106–29.
(32) Fisanick, W. The Chemical Abstracts Service Generic Chemical (Markush) Structure Storage and Retrieval Capability. 1. Basic Concepts. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 145–54.
(33) Fisanick, W. Storage and Retrieval of Generic Chemical Structure Representations. U.S. Patent 4,642,762, Feb 10, 1987.
(34) Ebe, T.; Sanderson, K. A.; Wilson, P. S. The Chemical Abstracts Service Generic CHemical (Markush) Structure Storage and Retrieval Capability. The MARPAT File. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 31–6.
(35) Fisanick, W.; Cross, K. P.; Rusinko, A., III. Similarity Searching of CAS Registry Substances. 1. Molecular Property and Generic Atom Triangle Geometric Searching. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 664–74.
(36) ISIS is an integrated scientific information system available from Molecular Design Limited, 2132 Farallon Dr., San Leandro, CA 94577. MACCS-3D is a module for 3D substructure search and retrieval.
(37) SYBYL/3DB UNITY is a software suite for 2D and 3D substructure search and retrieval available from Tripos Associates, Inc., 1699 S. Hanley Rd., Suite 303, Stl Louis, MO, 63144-2913.
(38) SYBYL is a molecular modeling software suite available from Tripos Associates, Inc., 1699 S. Hanley Rd., Suite 303, St. Louis, MO 63144-2913.
(39) ORACLE is relational database management system available from the Oracle Corp., 20 Davis Dr., Belmont, CA 94002.
(40) CAST-3D is available from Marketing Services, Chemical Abstracts Service, 2540 Olentangy River Rd., P.O. Box 3012, Columbus, OH 43210.
(41) Weinstein, H.; Maayani, S.; Srebrenik, S.; Cohen, S.; Sokolovshy, M. Psychotomimetic Drugs as Anticholinergic Agents. *Mol. Pharmacol.* **1973**, *9*, 820–33.
(42) Bartlett, P. A.; Shea, G. T.; Telfer, S. J.; Waterman, S. CAVEAT: A Program to Facilitate the Structure-Derived Design of Biologically Active Molecules. In *Molecular Recognition: Chemical and Biochemical Problems*: Roberts, S., Ed.; The Royal Society of Chemistry: London, 1989.
(43) Ullmann, J. R. An Algorithm for Subgraph Isomorphism. *J. Assoc. Comput. Mach.* **1976**, *23* (1), 31–42.
(44) Cheng, J. K.; Huang, T. S. A Subgraph Isomorphism Algorithm Using Resolution. *Pattern Recognit.* **1981**, *13* (5), 371–79.
(45) Cross, K. P.; Fisanick, W.; Rusinko, A., III. Atom-by-Atom Integrated 2D/3D Search Routines for Searching CAS Registry Substances, 202nd National Meeting of the American Chemical Society, New York, NY, Aug 1991.
(46) Sippl, M. J.; Stegbuchner, H. Superimposition of Three-Dimensional Objects: A Fast and Numerically Stable Algorithm for the Calculation of the Matrix of Optimal Rotation. *Comput. Chem.* **1991**, *15* (1), 73–8.
(47) Randic, M.; Jerman-Blazic, B.; Trinajstic, N. Development of 3-Dimensional Molecular Descriptors. *Comput. Chem.* **1990**, *14* (3), 237–46.
(48) Baldwin, J. J.; Ponticello, G. S.; et al. Thienothiopyran-2-sulfonamides: Novel Topically Active Carbonic Anhydrase Inhibitors for the Treatment of Glaucoma. *J. Med. Chem.* **1989**, *32*, 2510–3.
(49) Graham, S. L.; Hoffman, J. M.; et al. Topically Active Carbonic Anhydrase Inhibitors. 3. Benzofuran- and Indole-2-sulfonamides. *J. Med. Chem.* **1990**, *33*, 749–54.
(50) Eriksson, A. E.; Jones, T. A.; Liljas, A. Refined Structure of Human Carbonic Anhydrase II at 2.0A Resolution. *PROTEINS: Struct., Funct. and Genet.* **1988**, *4*, 274–82. Atomic coordinates available from the Brookhaven Protein Data Bank.
(51) Kannan, K. K.; Vaara, I.; Notstand, B.; Lovgren, S.; Borell, A.; Fridborg, K.; Petef, M. Structure and Function of Carbonic Anhydrase: Comparative Studies of Sulphonamide Binding to Hyman Erythrocyte Carbonic Anhydrases B and C. In *Drug Action at the Molecular Level*; Roberts, G. C. K., Ed.; University Park Press: Baltimore, MD, 1977.
(52) Viswanadhan, V. V.; Ghose, A. K.; Revankar, G. R.; Robins, R. Atomic Physico-Chemical Parameters for Three Dimensional Structure Directed Quantitative Structure–Activity Relationships. *J. Chem. Inf. Comp. Sci.* **1989**, *29*, 163–72.