Similarity Searching on CAS Registry Substances. 1. Global Molecular Property and Generic Atom Triangle Geometric Searching†

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

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

Received June 30, 1992

Chemical Abstracts Service (CAS) is exploring approaches for searching on 3D and related molecular property data for CAS Registry substances. This searching includes "fuzzy-match" similarity searching. As the first part of this effort, sample files have been created which contain 2D and 3D structure data and molecular property data. The 3D structure data are derived from the 3D coordinates that have been generated by the CONCORD program. The molecular property data such as partial atom charges, ionization potentials, and van der Waals volumes have been derived from the corresponding 2D and/or 3D data via computational chemistry programs. Experimental software is being developed to identify, analyze, and search various characteristics of 2D, 3D, and molecular property data for portions of the substance and/or the entire substance. This paper will discuss the general design of the test system and the analysis and searching of several data characteristics. Preliminary results indicate that fuzzy-match searching on global molecular property features appears to detect chemical or isosteric similarity and that fuzzy-match searching on generic atom triangle geometric features provides a significant amount of shape and size similarity.

I. INTRODUCTION

The searching of chemical substance data is a very important concept with respect to the access of information on chemical substances. Chemical Abstracts Service (CAS) has been interested in computer-based substance searching since the late 1960s when we developed a batch substructure search system based on the structure (or topology) of a specific chemical substance as represented in a connection table.[1] During the early 1970s, efforts were focused on the development of techniques for searching the nomenclature representations of specific chemical substances.[2,3] In 1977, work began on topological-based techniques for online, substructure searching of the specific substances in the CAS Chemical Registry; this led to the online Registry File substructure search system, which is available today through STN International.[4-6]

In the early 1980s, research began on the use of topologically based techniques for the handling of generic chemical (Markush) substances. This research led in 1985 to the capability for formulating generic queries for searching specific substances on the Registry File on STN, and later to the capability for storing and searching generic chemical substances.[7-9] The initial application of CAS's generic substance handling is a search service on Markush structures from patents that is available as the MARPAT File on STN.[10]

Recently, CAS has been exploring several additional capabilities which could lead to a significant expansion of the scope of our handling of substances.[11-13] Our research is based on an overall view that substance handling should involve three fundamental substance data types and three fundamental search types for a total of nine individual capabilities.[11]

The three fundamental substance data types are *2D structure*, *3D structure*, and *molecular property* data (chemical nomenclature data is also useful as a surrogate for 2D structure data). 2D structure data is available for over 11 million substances in the CAS Registry File and for approximately 100 thousand generic (Markush) structures in the MARPAT File. CAS currently has 3D coordinates for over 4.6 million Registry substances. These coordinates were generated by the CONCORD program[14] and currently are available only on a retrieval basis via STN. (One of the objectives of on-going research activities is to ascertain the feasibility of directly searching this large file of 3D structures.) The initial molecular property data that we are experimenting with have been computer-generated using computational chemistry programs such as the semiempirical, molecular orbital package, MOPAC.[15]

When searching the data types, *global* and *local* substance features are distinguished. A global feature applies to the entire molecule such as an ionization potential, a dipole moment, or a topological index. A local feature applies to just a portion of a molecule such as a 2D structural fragment, the partial charge on an atom, or the interatomic distance between a pair of atoms. Several local features may also be grouped into a feature set such as a connected set of structural fragments. Features may be *specific* or *generic*. For example, the number of bonds in a substance is a generic feature, whereas an augmented atom fragment (a central atom and its nearest neighbors) with specified atoms and bonds is a specific feature.

Three search types are distinguished for matching query features with file substance features. These types are illustrated in Figure 1 using "connected" local features (LF) and global features (GF). An example of a connected pair of local features is a "chloro" bonded to a "benzene" ring. The shaded area of the circles in the column on the far right represents the common features, i.e., the "overlap", among the query and file substance features. In an exact or *identity-match* search (Figure 1a), *all* the query features are located in a file substance, and the file substance has no additional features. An example of this is a 2D full (exact) structure search. It should be noted that an identity-match search on 3D structure and molecular property data is probably not very useful since the exact numeric values (e.g., interatomic distances) usually correspond to the conformational model generated and can vary significantly from model to model. Thus, some "fuzziness" in searching these data types is

† Presented in part at the 1992 Beilstein Symposium on Similarity in Organic Chemistry held in Bozen (Bolzano), Italy, May 25–29, 1992.

0095-2338/92/1632-0664$03.00/0

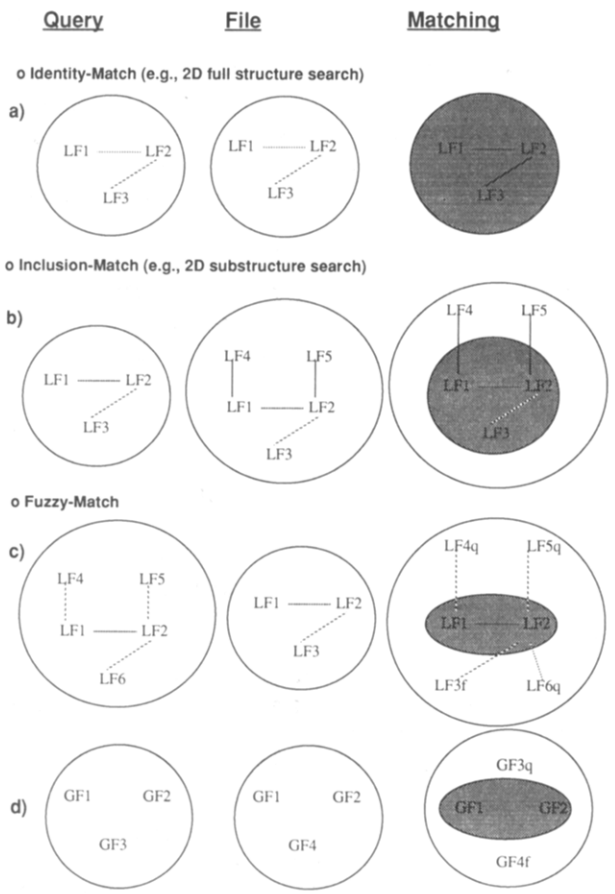


Figure 1. Search-type matching criteria.

probably desired in most cases via numeric range searching.

In an *inclusion-match* search (Figure 1b), *all* the query features are located in the file substance, but the file substance may have additional features or vice versa. An example of this type of search is a 2D substructure search. The case where the file structure is included in the query structure has been referred to as a superstructure search.

The third search type is *fuzzy-match* search. Typically, in this type of search, only *some* of the query features need be located in the file substance for retrieval to occur. A similarity metric is usually calculated which measures the overlap of the query and file substance features and ranks the retrieval. In some cases it is desirable to rank the entire database relative to the query. Fuzzy-match searching is referred to by many as "similarity" (or "dissimilarity") searching.[16,17] Figure 1c illustrates a fuzzy-match search involving local features; Figure 1d does the same for global features.

A basic goal of CAS's current research in substance handling is to determine the feasibility of capabilities for executing the three search types against the three data types, except for 2D structure identity and inclusion-match capabilities which are already available. This research includes an investigation of an integrated search capability on large files of Registry substances. This integration is with respect to both data types and search types. An example of this integration would be a query which specifies that a 2D structural fragment be fuzzy-matched, that a 3D structural fragment be inclusion-matched, and that a global, molecular property value be matched within a specified tolerance.

As a prototype for these additional capabilities, CAS has been implementing various algorithms in an experimental system for similarity and 3D searching in a client–server architecture. Thus far, a variety of techniques for query-framing and refinement searching have been implemented, including an integrated, 2D/3D atom-by-atom search[12] and a numeric search capability for molecular property data. A variety of experimental substance databases are being used, including the CAS 3D Structure Templates (CAST-3D) database of approximately 370 000 *rigid* and *semirigid* Registry substances with 3D coordinates which contains a significant number of ring system substances.[11,13] There is also a database of computer-generated molecular properties for 60 000 Registry substances. Experimental fuzzy-match capabilities have been developed and used to obtain the results described in this paper, but, as yet, they have not been added to the prototype system.

The focus of this paper is on the relationships among a set of integrated 2D, 3D, and molecular property features and the use of fuzzy-match techniques for searching these integrated features. More specifically, some results of fuzzy-match similarity searching on global molecular property and generic triangle geometric features are reported. We expect that some of these results will lead to the incorporation of new capabilities into the developing experimental similarity and 3D searching system mentioned above.

Fuzzy-match similarity search procedures for 2D structure data are fairly well developed.[16,17] Typically, the techniques are *fragment-based* and involve the use of predetermined fragments such as substructure search screen fragments. More compute-intensive techniques which involve the in situ determination of common substructures between the target structure and the file structures are also used. Also, another possible technique is one that utilizes a generic structure as a framework for a similarity comparison. We are currently investigating the use of a particular type of generic structure[8] as part of such a technique.

Pepperrell et al. have been exploring various techniques for fuzzy-match searching of 3D structures using interatomic distances.[18,19] They found that the most cost-effective technique is an atom-mapping method that identifies pairs of atoms between the target substance and file substance which have neighboring atoms at approximately the same distances.

Fuzzy-match similarity searching on molecular property data is, to our knowledge, a new concept. We have recently reported on a statistical-based method for performing such searching.[11] This technique is described in more detail later.

The purpose of fuzzy-match searching is to collect together a set of substances that are "similar" to a target substance. The use of different data types and a variety of generic and specific feature classes within each data type allows a variety of substance similarity "views". For example, the use of 2D structural features such as augmented atoms, bond counts, and topological indices may indicate *structural similarity*. Likewise, inherent in an appropriate set of 3D structure features there is at least the potential for *shape and size similarity*. An appropriate set of molecular property features may indicate *chemical or isosteric similarity*, i.e., substances with similar chemical and physical properties, especially those that may be structurally dissimilar. It is important to note that some of the molecular property data are based on the 3D structure and that CONCORD generates one, low-energy 3D structure or conformer per substance. It is known that the values of most features will vary with different conformers of a substance. Similarly, 3D structural features may also vary from conformer to conformer. The fuzziness in the searching should allow for some of the differences among conformers. However, the results of fuzzy-match searching on molecular properties and 3D structural features are viewed as being

illustrative and the source of new ideas, i.e., searching based on a single conformer is not exhaustive.

The ability to specify different similarity types or type combinations for a target substance is particularly important in locating suitable ligands that might interact with a biological receptor. Shape, electrostatic, H-bonding, and hydrophobicity factors are important for such interactions. In some cases, a more detailed comparison of shape, electrostatic potential, etc. will be needed, and several techniques have been developed for this purpose.[17] Thus, fragment-based searching on features could be an initial or *screening* step for such time-consuming refinement procedures.

Ideally, a system for fragment-based, fuzzy-match searching on 2D, 3D, and molecular property data should have the following characteristics:

1. appropriate feature classes for the desired similarity types
2. features within a class should be *orthogonal* to the extent possible
3. appropriate fuzzy-match methods for the various feature classes, including a feature weighting and answer threshold mechanisms
4. automatic profiling for major similarity types via the input of a CAS Registry Number or a 2D structure
5. ability for a user to specify any combination of features or feature classes to be used as a profile for a target substance

Using these characteristics as a guideline, we have developed a suite of software and files for modeling fragment-based, fuzzy-match searching which we call a Substance Similarity Search Modeller (SSSM). The modeller contains search, retrieval, and analysis software in the Statistical Analysis System (SAS)[20] and an experimental database with over 3000 features for approximately 6000 substances. Our overall strategy has been to examine a large number of substance features with the modeller and then to ascertain pertinent information such as which features are complementary and which ones are redundant with respect to a similarity type, what is the relationship between global and local features, etc. We plan to implement desirable subsets of features and fuzzy-match techniques for the larger databases in the prototype system for further evaluation and to obtain user feedback.

II. CONTENT OF EXPERIMENTAL DATABASE

The database used in our testing was derived from a systematic sample of the 4.6 million Registry substances which currently have CONCORD-generated 3D coordinates. Up to 3172 features currently may be present for each substance. There are a maximum of 161 molecular property features, 663 2D features, 2230 2D/3D features, and 118 3D features. 2D/3D features are those that contain specific nodes as well as geometric information. For example, the interatomic distance between a carbon and oxygen atom pair is a 2D/3D feature. 3D features consist solely of geometric information such as an interatomic distance or the perimeter of a triangle of atoms.

The typical generation procedure for the features involves first generating "raw" fragments or values from the data types. For example, approximately 15 million atom triangles were generated from the 3D structures of the database, i.e., approximately 2500 triangles per substance. The raw data is then "reduced" into the feature definitions, including any "binning", via feature generation software. The features values are integer or floating point numbers which are used by the search system software. It is important to note that these feature definitions are more compact than typical 2D or 3D substructure screen fragments. For example, a count of carbons is a single feature, but for the STN Registry File, 14 screens are used to handle the carbon count. The feature set would probably correspond to a screen set which has a total number of screens perhaps 10–20 times the number of features.

The following are a brief description of the various classes of features. Where appropriate, CONCORD-generated 3D structures were used as models for the calculation of features.

1. Global (MP) Molecular Properties

   These are calculated properties that were generated via MOPAC, SAVOL, and CAS software. Some examples of global molecular properties are ionization dipole moments, heat of formation, and molar refractivity. Specific numeric values for the properties are used.

2. Local (MP) Molecular Properties

   These properties were generated via MOPAC. Some examples of local molecular properties are atomic electron densities and eigenvalues for the molecular orbitals. The values are "binned", and bin counts are used.

3. Local (2D) Path Length Features

   These are occurrence counts for 1–31 path lengths with Any- - -Any, Carbon- - -Carbon, Carbon- - -Hetero, and Hetero- - -Hetero terminal atoms. The path length identifies the shortest path between the two nodes.

4. Local (2D/3D) Path Length Features

   These are interatomic distance sums for 1–31 path lengths with Any- - -Any, Carbon- - -Carbon, Carbon- - -Hetero, and Hetero- - -Hetero terminal atoms, i.e., the distance between the terminal atoms. The path length is for the shortest path between the two nodes.

5. Global Topological (2D) Index and Related Features

   These are global topological index features generated primarily via MOLCONN2 software.[21] Some examples are the number of edges (bonds), the κ indices,[22] and the Shannon index.[23]

6. Global (2D) Flexibility Indices

   These are mean/sum indices, which include CAS's Local Simple and Local Simple Normalized indices along with Kier's π index.[11,24]

7. Local (2D) Flexibility Indices

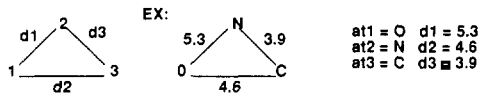
   These are occurrence counts of Local Simple (LS) and Local Simple Normalized (LSN) indices.[11] LS has 62 bins, for values from 1 to 31 linearly incremented by 0.5 unit. LSN has 31 bins, for values 0–0.725 linearly incremented by 0.025 unit.

8. Local (2D/3D) Flexibility Indices
   These are terminal atom, interatomic distance sums for CAS LS and LSN indices. LS has 62 bins, for values from 1 to 31 linearly incremented by 0.5 unit. LSN will have 31 bins, for values 0–0.725 linearly incremented by 0.025 unit.
9. Global/Local (2D) Ring Analysis Data Features
   These are mostly local features based on ring analysis data for ring systems contained in a substance. Included are the total number of ring systems per substance and, per ring system, the number of rings, size of rings, the ring analysis, and component ring analysis. For example, for quinoline (a mono nitrogen heterocycle consisting of two fused six-membered rings), the ring analysis is C_5N–C_6, the component ring analyses are C_5N and C_6, the number of rings is 2, and the size of the rings is 6,6.
10. Local (2D) Molecular Formula Features
    These are the occurrence counts for H, C, N, O, F, Si, P, S, Cl, Br, and I.
11. Local (2D) Generic Features
    These are bond counts, degrees of connectivity counts, etc.
12. Local (3D) Atom-Pair Distance Bin Counts
    These are distance bin counts for Any- - -Any, Carbon- - -Carbon, Carbon- - -Hetero, and Hetero- - -Hetero atom pairs. The Lederle binning formula is used, i.e., bin number = 5 arctan $[(D - 3.0)/2] + 20$, where D is the interatomic distance.[25]
13. Local, (3D) "3-Bonded Atoms" Angle Bin Counts
    These are angle bin counts for bonded atom triplets involving Any, Carbon, and Hetero atoms. The bins have been selected to encompass the common atomic hybrid angles (i.e., 109.5, 120, and 180 degrees).
14. Local, (3D) "4-Bonded Atoms" Angle Bin Counts
    These are dihedral angle bin counts for bonded atom quartets involving Any, Carbon, and Hetero atoms. The bins are equifrequency based on distribution statistics, i.e., a frequency histogram was used to select bins containing an approximately equal number of angle fragments.
15. Local, (3D) "3 Atoms and 1 Bond Vector" Angle Bin Counts
    These are angle bin counts for atom triplets involving Any, Carbon, Hetero, and Hydrogen atoms. Two atoms are bonded while the third atom can be any atom in the structure. Only one hydrogen is permitted. The bins are equifrequency based on distribution statistics.
16. Local, (3D) "4 Atoms and 2 Bond Vectors" Angle Bin Counts
    These are dihedral angle bin counts for atom quartets involving Any, Carbon, Hetero, and Hydrogen atoms. The dihedral angle between two bond vectors was computed. The bond vectors are not connected. The bins are equifrequency based on distribution statistics.
17. Local, (2D/3D) Atom Triangles, and (3D) Related Metric Bin Counts
    These are triangle atoms and related metrics such as the area and perimeter of the triangle. The atoms are Any, Carbon, Nitrogen, Oxygen, Hetero, H-Donor, H-Acceptor, and a Hetero other than Nitrogen or Oxygen. The bins are equifrequency based on distribution statistics.
18. Local, (2D/3D) Atom Triangle "3-Slot" Bin Counts
    These are the first two ordered atoms in a triangle and their binned interatomic distance. The bins are equifrequency based on distribution statistics.
19. Local, (2D/3D) Atom Triangle "5-Slot" Bin Counts
    These are the first two ordered atoms in a triangle and the ordered set of three binned interatomic distances in the triangle. The bins are equifrequency based on distribution statistics.
20. Generic Key (2D, 3D, MP) Features
    These are generic keys built by normalizing the values of the features in each feature class and certain groups of classes plus a "total" substance key built from the individual classes.

In addition to the above classes, we are in the process of adding an augmented atom class of features (2D) to the database. It is expected that this class of features will increase the total number of substance features by an additional 500–1000.

The experimental results discussed in the subsequent sections involve the global molecular property classes and several of the triangle-based feature classes. The following are brief descriptions of the molecular properties selected for testing. Except where indicated, the data were generated by MOPAC (1 SCF, i.e., not a complete optimization) using CONCORD 3D coordinates as input.

1. Heat of Formation (ht_form)—the heat of formation is relative to elements in their standard state.
2. Total Energy (tot_eng)—the total energy is the sum of the electronic and nuclear terms.
3. Ionization Potential (ion_pot)—the ionization potential is energy needed to remove an electron from the system and can be approximated as the negative of the highest occupied or the highest partially occupied molecular orbital.
4. Dipole Moment (dsum_t)—the dipole is the magnitude of the dipole vector computed from the vector sum of the x, y, and z dipoles.
5. Lowest Unoccupied Molecular Orbital (lumo)—the LUMO is the "$n + 1$" eigenvalue where "n" is the number of filled orbitals.
6. Difference between HOMO and LUMO Values (del_h_l).
7. Charge Mean (chg_mean)—the mean of all partial atomic charges.
8. Charge Standard Deviation (chg_std)—the deviation from the mean of all partial atomic charges.
9. Electron Density Mean (ed_mean)—the mean of all atomic electron densities.



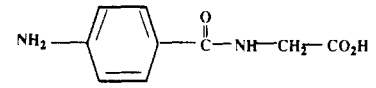
• "3-Slot": at1_at2_d1 Counts (Equi-Freq. Bins) --	110
• "5-Slot" : at1_at2_d1_d2_d3 Counts (Calc. Bins) --	1586
• Triangle Atoms: at1 & at2 & at3 Counts --	16
• Triangle d1, d2 & d3 Edges (Equi-Freq. Bins) --	30
• Triangle Height Counts (Equi-Freq. Bins) --	10
• Triangle Area Counts (Equi-Freq. Bins) --	10
• Triangle Perimeter Counts (Equi-Freq. Bins) --	10
• Triangle Inscribed Radius Counts (Equi-Freq. Bins) --	10
• Triangle Perimeter/Largest Leg Counts (Equi-Freq. Bins)	10
	1792

Figure 2. Triangle features on the experimental database.

10. Electron Density Standard Deviation (ed_std)—the deviation from the mean of all electron densities.
11. Electron Density Maximum (ed_max)—the maximum value of all the atomic electron densities.
12. Electron Density Minimum (ed_min)—the minimum value of all the atomic electron densities.
13. Octanol/Water Partition Coefficient (logP)—calculated using the Ghose and Crippen method.[26]
14. van der Waals Volume (vol)—calculated using the SAVOL program.[27]
15. Number of Filled Orbitals (fil_orb).
16. Highest Occupied Molecular Orbital (homo)—the HOMO is the "n" eigenvalue where "n" is the number of filled orbitals and is considered to be the energy of that molecular orbital.
17. Charge Maximum (chg_max)—the maximum value of all the partial atomic charges.
18. Charge Minimum (chg_min)—the minimum value of all the partial atomic charges.
19. Molar Refractivity (mr)—calculated using the Ghose and Crippen method.[26]
20. van der Waals Surface Area (sa)—calculated using the SAVOL program.[27]

Calculation of Pearson correlation coefficients between each of the 20 features reduced the set from 20 to 14 (the first 14 shown above). If a feature correlated with another feature with a coefficient greater than 0.9, then the second feature was considered redundant and was discarded. The features retained depend, of course, on the order in which they were examined. For example, since the correlation coefficient between the VDW volume and surface area was greater than 0.99, only one of these terms was kept.[28] The VDW volume was retained since it was encountered first.

Features based on atom triangles (2D/3D and 3D) are illustrated in Figure 2. The different triangle classes and the number of features within the classes are shown. As mentioned above, approximately 2500 atom triangles are generated for each database substance from the 3D structure. The atoms in the triangle are ordered based on atomic number. Thus, in Figure 2, O is 1, N is 2, and C is 3. The distance ordering is d1 between atoms 1 and 2; d2 between 1 and 3; and d3 between 2 and 3. If there are equivalent atoms, then the distances are ordered by size. The most specific feature class is the "5-slot" counts which encodes the first two atoms and



61-78-9
p-Aminohippuric Acid

GS Flexibility Index	Heat of Formation	Ionization Potential	Logp	Any---Any (PL=3) Dist. Sum	2D Wiener Index
3.11	-88.996 kcal	8.589 ev	-0.11	52.58A°	33

C6 Ring Analysis	O-Atom Count	H-Atom Charge(B11)	Non-Bonded Angle H-Ht---C(B2)	Non-Bonded Dihedral Angle C-Ht---C-C(B5)
1	3	4 (B11=.162-.280)	8 (B2=87-105°)	31 (B5=111-180°)

Triangle Height(B8)	Triangle Atoms (Hd,Ht,C)	Triangle 3-slot Ha---C Dist. (B5)	Triangle 5-slot OCB1B1B1	Generic Key
39	184	12	1	0.069
(B8=2.20-2.65A°)		(B5-4.3-4.8A°)	(B1=0.0-2.1A°)	

Figure 3. Selected features for *p*-aminohippuric acid.

the three distances. These features were designed primarily for inclusion-match (substructure) searching since five out the six components necessary to uniquely identify a triangle are encoded (the third atom is typically a carbon, and thus, only a small amount of additional discrimination would be expected by encoding it).[11] Only the last six classes in Figure 2, containing 80 3D features, were used in the experiments described in this paper. The set reduces to just 29 features when redundant features were discarded.

Figure 3 illustrates some selected features for the substance, *p*-aminohippuric acid. This substance has a (nonzero) feature vector of 1088 features. The GS Flexibility Index (3.11) is a topological index developed by CAS as an approximate measure of conformational flexibility.[11] The Any- -Any (PL = 3) feature is the sum (52.58 Å) of the interatomic distances of all atom pairs that are separated by a shortest path of three. The H-Atom Charge feature indicates a count of four H-atoms in "bin" 11, i.e., with a charge between 0.162 and 0.280. The Non-Bonded Angle ("3 Atoms and 1 Bond Vector") feature indicates a count of eight angles between 87 and 105° (bin 2) for an atom triplet in which a hydrogen is bonded to a hetero atom (Ht), which in turn has a nonbonded distance to a carbon atom. The Non-Bonded Dihedral Angle ("4 Atoms and 2 Bond Vectors") feature indicates a count of 31 angles between 111 and 180° (bin 5) for an atom quartet in which the central nonbonded atoms are hetero (bonded to carbon) and a carbon (bonded to a carbon). The triangle height feature indicates a count of 39 triangles with a height (longest edge to the opposite node) between 2.20 and 2.65 Å (bin 8). The Triangle Atoms feature indicates a count of 184 triangles with the atoms H-Donor (Hd), hetero, and carbon. The Triangle 3-Slot feature indicates a count of 12 where the ordered atom no. 1 is H-Acceptor (Ha), atom no. 2 is carbon, and distance no. 1 is between 4.3 and 4.8 Å (bin 5). The Triangle 5-Slot feature indicates a count of one and atom no. 1 is oxygen, atom no. 2 is carbon, and distance no. 1, distance no. 2, and distance no. 3 are between 0.0 and 2.1 Å (bin 1). The Generic Key is a value that encodes the entire set of 2D, 3D, and MP features in the substance. It is built by taking the mean of all the generic keys for the individual classes.

III. SUBSTANCE SIMILARITY SEARCH MODELER

As mentioned above, the Substance Similarity Search Modeller (SSSM) is a collection of search, retrieval, and analysis software in the Statistical Analysis System (SAS).

```
                    19MAY92
        SUBSTANCE SIMILARITY SEARCH MODELLER

   FUZZY-MATCH SIMILARITY SEARCH PROFILE PARAMETERS
        6K INTEGRATED 2D/3D/MP DATABASE

TITLE=FRS8Q1A_MP_GLOBAL_FULL_LESS_HT_FORM
QRY_REG=6306736
SCORING MECHANSIM = STATISTICAL METHOD/VARIABLE RANGE
   FACTORS = 1 , 0.75 , 0.5 , 0.25
NO. OF SCREENS =0     SCREEN THRESHOLD =0 %
NO. OF FEATURES =20   FEATURE THRESHOLD =80 %

                    FEATURES

      Id            Value       Weight

ht_form =           -3.97          0
tot_eng =        -2754.734         1
ion_pot =            9.512         1
dsum_t =             4.379         1
lumo =              -0.108         1
del_h_l =           -9.404         1
chg_mean =              0          1
chg_std =            0.204         1
ed_mean =            2.929         1
ed_std =             1.933         1
ed_min =             0.72          1
ed_max =             6.41          1
logp =               1.46          1
vol =              197.684         1
fill_orb =          41             1
homo =             -9.51227        1
chg_max =            0.389         1
chg_min =           -0.4102        1
sa =               240.469         1
mr =                59.84          1
```

Figure 4. Search parameters for a sample query.

It also contains supporting Unix search, answer, display, and statistic files. The SSSM interfaces with the above-mentioned prototype system via answer files of Registry Numbers. Both inclusion and fuzzy-match similarity searching are supported. The search file is a large matrix of feature columns and substance rows. The searching is performed directly on the numerical values of the features. Features can be used as "screens" for other features, i.e., these other features will not be searched unless the substance passes the specified feature screens. In fuzzy-match searching, the system will automatically build a search profile given a Registry Number of the target substance, and the identifiers of the features to be used in the searching. (Currently, the target substance must be on the database to accomplish this automatic profiling.)

Figure 4 illustrates a display of the search parameters for a sample fuzzy-match search. The Registry Number of the target substance is 6306-73-6 (1-benzylthymine). The scoring mechanism selected is a "statistical method/variable range" which is discussed in more detail below. A total of 20 features are specified with no feature screens. Only those substances with a score greater than the 80% similarity threshold will be retrieved. The Id and Weight columns echo the user input. The Value column is derived automatically in a preliminary search of the data base for Registry Number 6306-73-6 and is the basis for determining the search ranges for the features.

There are currently two basic scoring methods being used in fuzzy-match searching. These are a "presence/absence" method and a "statistical" method. In the presence/absence method, each file substance feature is a "hit" relative to the corresponding query feature, if it has a value greater than zero. A simple percent overlap of the query features to file substance features can be specified as a scoring mechanism. Alternatively, the calculation of a Tanimoto coefficient can be requested.[16]

The statistical method was the method used in the experiments described below. There are "static range" and "variable range" versions of this method. The calculation of a static range is illustrated in Figure 5. The target substance is *N*,*N*-dimethyl-1,3-propanediamine (109-55-7), and the illustrative feature is the ionization potential. The standard deviation from the mean of all substance ionization potentials is 1.00. A total of four search ranges are used for each target feature. A range is computed by taking the target feature

```
                                  1,3-Propanediamine, Database
                                  N,N-dimethyl-     Std. Dev
Ionization Potential (IP)              9.58ev          1.00ev

Range # 1 (8.58-10.58) = 9.58 ± (1.00 * 1)
Range # 2 (8.83-10.33) = 9.58 ± (1.00*.75)
Range # 3 (9.08-10.08) = 9.58 ± (1.00*.50)
Range # 4 (9.33-9.83)  = 9.58 ± (1.00*.25)
                          ↑      ↑    ↑    ↑
                        Range  Value Std  Factor

IF IP <8.58-10.58> THEN IP_SCORE=IP_SCORE+1  /*1.00std */
IF IP <8.83-10.33> THEN IP_SCORE=IP_SCORE+1  /*0.75std */
IF IP <9.08-10.08> THEN IP_SCORE=IP_SCORE+1  /*0.50std */
IF IP <9.33- 9.83> THEN IP_SCORE=IP_SCORE+1  /*0.25std */

        IP_SCORE=IP_WEIGHT*IP_SCORE
                        o
                        o
   TOTAL_SCORE=IP_SCORE+Other feature scores
```

Figure 5. Illustration of the similarity calculation for the ionization potential of *N*,*N*-dimethyl-1,3-propanediamine.

```
                    19MAY92
        SUBSTANCE SIMILARITY SEARCH MODELLER
        FUZZY-MATCH SIMILARITY SEARCH RESULTS
        6K INTEGRATED 2D/3D/MP DATABASE

Query Reg_no=6306736
Profile Title =FRS8Q1A_MP_GLOBAL_FULL_LESS_HT_FORM
Zero-Zero Match Wt.=4
Scoring Mechanism = Statistical Method/VARIABLE RANGE
Factors = 1 , 0.75 , 0.5 , 0.25
No. of Screens =0
Max. Screen Score=0
No. of Features =20
Max. Feature Score=76
No. of Non-Zero Query Features =19

 Substance Answers at:

  0  Percent Screen Threshold
  80  Percent Feature Threshold

T_S Pct  Reg     No. Feature Scores=>=>=>=>=>=>=>=>=>=>Feature Data Values

76  100  6306736  1  ht_form=0  tot_eng=4  |||  ht_form=  -3.97      tot_eng=  -2754.734
75  98.7 7269047  2  ht_form=0  tot_eng=4  |||  ht_form=  29.422     tot_eng=  -2818.712
74  97.4 4464920  3  ht_form=0  tot_eng=4  |||  ht_form=- 42.683     tot_eng=  -2690.987
73  96.1 1781346  4  ht_form=0  tot_eng=4  |||  ht_form=  -1.521     tot_eng=  -3073.31
72  94.7 4474515  5  ht_form=0  tot_eng=4 oo|||  ht_form= -16.061     tot_eng=  -2852.977 oo
72  94.7 4702798  6  ht_form=0  tot_eng=4  |||  ht_form=  50.546     tot_eng=  -2596.833
72  94.7 5013901  7  ht_form=0  tot_eng=4  |||  ht_form=161.965      tot_eng=  -2846.098
72  94.7 5628546  8  ht_form=0  tot_eng=4  |||  ht_form=-92.748      tot_eng=  -2755.388
72  94.7 6981642  9  ht_form=0  tot_eng=4  |||  ht_form=-62.024      tot_eng=  -2874.104
71  93.4 713906  10  ht_form=0  tot_eng=3  |||  ht_form= 19.459      tot_eng=  -2361.433
```

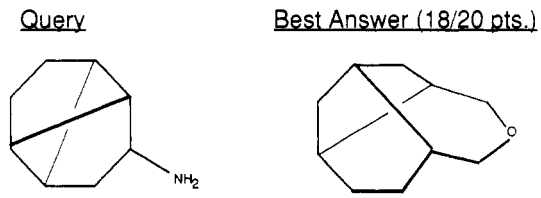
Figure 6. Results display from a sample fuzzy-match similarity search.

value and adding and subtracting the database standard deviation times an inclusion factor. The inclusion factors can be specified by the users. In this case, the factors are 1.0, 0.75, 0.50, and 0.25.

The difference between the static range and variable range version is that with the static range the value of one standard deviation is multiplied by the inclusion factor, whereas with the variable range the number of standard deviations will vary for the different ranges. This number is based on the computation of the "normal score" for the target feature. Essentially, the further the target feature value is from the mean of the feature for all the database substances, then the wider the search range. A lookup table is used to retrieve the number of standard deviations to be used in the calculation via the value of the normal score.

The calculation of a score for a feature and the total substance score are illustrated by the pseudo-code in the bottom part of Figure 5. A score of one is added to the feature score if the file substance feature value falls within a search range. Thus, a feature score can range from zero (not included in any of the ranges) to four (included in all the ranges). The score is then multiplied by a user-supplied weight. The total substance score is the sum of all the feature scores.

One type of SSSM retrieval display is illustrated in Figure 6. This retrieval is for the search parameters given in Figure 4. The total score (T_S) and the percent score (PCT) are given along with, as a user option, the feature scores and the file substance feature values. Note that the target substance is completely matched and returns a value of 100%.



CAS RN 15774-49-9		CAS RN 250-19-1
Ionization Potential	7.25ev	7.33ev
Dipole Moment	2.50Debyes	2.77Debyes
HOMO - LUMO	-7.73ev	-8.68ev
Mean Electron Density	2.27	2.38
Van der Waal's Volume	127.15Å^3	116.02Å^3

Figure 7. Target substance and "best" answer for a fuzzy-match search using five global molecular property features.

Figure 7 illustrates the best result of a molecular property search involving five features, using the static range version of the statistical method.

All the experiments described below used the variable range method. It should be noted that the frequency distributions for the global molecular properties are fairly normal. This is illustrated by the ionization potential distribution shown in Figure 8.

Perhaps the most significant characteristic of the SSSM is that essentially any number and combination of features can be expressed as an inclusion or fuzzy-match search. This allows for comparison of features and feature classes within a data type and between data types. Thus, for example, searches can be performed using an integrated set of 2D, 2D/3D, 3D, and MP features or any combination thereof. In addition to the comparison of features via fuzzy-match searching, we are also using the SSSM to model a set of screens for 3D substructure searching. This is done by defining a "logical" screen dictionary and by using the appropriate numeric range search on the features such that there is a mapping to the logical dictionary.

IV. EXPERIMENTS AND RESULTS

A test set of 20 target substances was selected as a systematic sample of the database substances. The sample consists of some typical organic substances. Some examples are illustrated in Figure 9 (see also the substance in Figure 3 and the targets in Figures 11 and 12). The substance shown at the bottom of Figure 9 has 36 heavy atoms and is the "largest" substance in the sample. This query set was used in both the molecular property and triangle feature searching.

Global Molecular Property Searching. Both the full (20 features) and reduced (14 features) sets were used in searching. The reduced set searches yielded a collective total of 896 hits, including the target substances, using an 80% similarity threshold and inclusion factors of 1.0, 0.75, 0.50, and 0.25. The mean percent similarity score was 84.1. The distribution of these answers in terms of percent scores is shown in Figure 10. The 100% similarity scores are the target substances. This distribution is what should be expected since there is a decrease in the number of substances as the percent similarity increases. The distribution would be expected to be essentially the same for large files of substances such as the full Registry file, except, of course, the size of the retrieval would be proportionally larger. Thus, for the full file one might expect several hundred answers at greater than say 98% similarity to the target substance.

The discrimatory power of the scoring method is reflected in the redundancy of the scores, i.e., the smaller the number of redundant scores, the greater the discriminatory power of the method. The number of redundant scores were determined for the answer set of reduced set queries with more than 20 answers (a total of 12 queries). The mean number of redundant scores was 14 in the first 20 answers and, thus, the scoring method is not very discriminating. However, the use of smaller inclusion factors can increase the discrimatory power somewhat. For example, a set of four reduced set queries were executed using the inclusion factors 0.4, 0.3, 0.2, and 0.1. The mean number of redundant scores dropped from 15 (inclusion factors 1.0, 0.75, 0.50, and 0.25) to 12.

Table I. Prediction of Global Molecular Properties for RN 6306-73-6

feature	reduced set predicted value (STD)	full set predicted value (STD)	feature value	database STD
dsum_t	3.42 (1.63)	3.41 (1.66)	4.38	2.43
ion_pot	9.23 (0.57)	9.29 (0.57)	9.51	1.00
logp	1.61 (1.35)	1.68 (1.37)	1.46	2.14
ht_form	−21.07 (69.54)	−20.61 (71.54)	−3.97	137.64
tot_eng	−2901.90 (601.50)	−2788.41 (376.57)	−2754.70	1320.16
lumo	−0.24 (0.51)	−0.24 (0.60)	−0.11	1.21
del_h_l	−9.04 (0.31)	−9.08 (0.74)	−9.40	1.49
chg_mean	0.00 (0.00)	0.00 (0.00)	0.00	0.01
chg_std	0.18 (0.02)	0.18 (0.03)	0.20	0.14
ed_mean	2.84 (0.23)	2.86 (0.27)	2.93	0.54
ed_std	1.96 (0.10)	1.97 (0.11)	1.93	0.20
ed_min	0.78 (0.05)	0.78 (0.05)	0.72	0.26
ed_max	6.49 (0.30)	6.49 (0.30)	6.41	0.59
vol	208.97 (41.49)	201.61 (25.36)	197.68	81.71

To determine whether or not our techniques produces statistically significant results, we built a "random" database containing the 14 molecular property features for the 6K substances and executed searchers for several queries. A total of 20K random, pairwise exchanges of values were performed on each feature column for the substances. (NOTE: The overall distributions of the database feature values are not affected by this processing.) For a set of 5 queries which had 488 hits at the 80% similarity threshold in the regular database, only one hit (at 80.4% similarity) was retrieved in the searches on the random database. Thus, the technique produces results that are significantly better than what can be obtained by chance.

The "leave one out" technique was used as a way of checking on the validity of the similarities. One of the queries was repeatedly executed 14 times. In each search, a different feature was "dropped out" (by assigning a feature weight of zero), and the value of this feature for the target substance was predicted from the retrieval results. The mean of the feature values for the most similar substances to the target substance was used as the predicted value. The procedure was also performed for the full set of features (20). The results are shown in Table I. All retrievals at greater than 80% similarity were used in the calculations. Considering the standard deviations (STD) from the mean for the feature for all database substances, the predictions were very good, i.e., similar substances have similar properties. Note that the full set and the reduced set predictions are very similar. We are currently examining property prediction using narrower percent similarity ranges, i.e., >95% 90–95%, 85–90%, etc.

Our goal in experimenting with global molecular properties was to determine if such features could provide for chemical or isosteric similarity and, hence, complement features used to describe structural (2D) similarity. Perhaps the most significant observation in molecular property search results is that many substances with high similarity scores are topologically (2D) somewhat different from the target substance. This observation is born out as the "best" answer shown in Figure 7. This structure has a similar basic ring

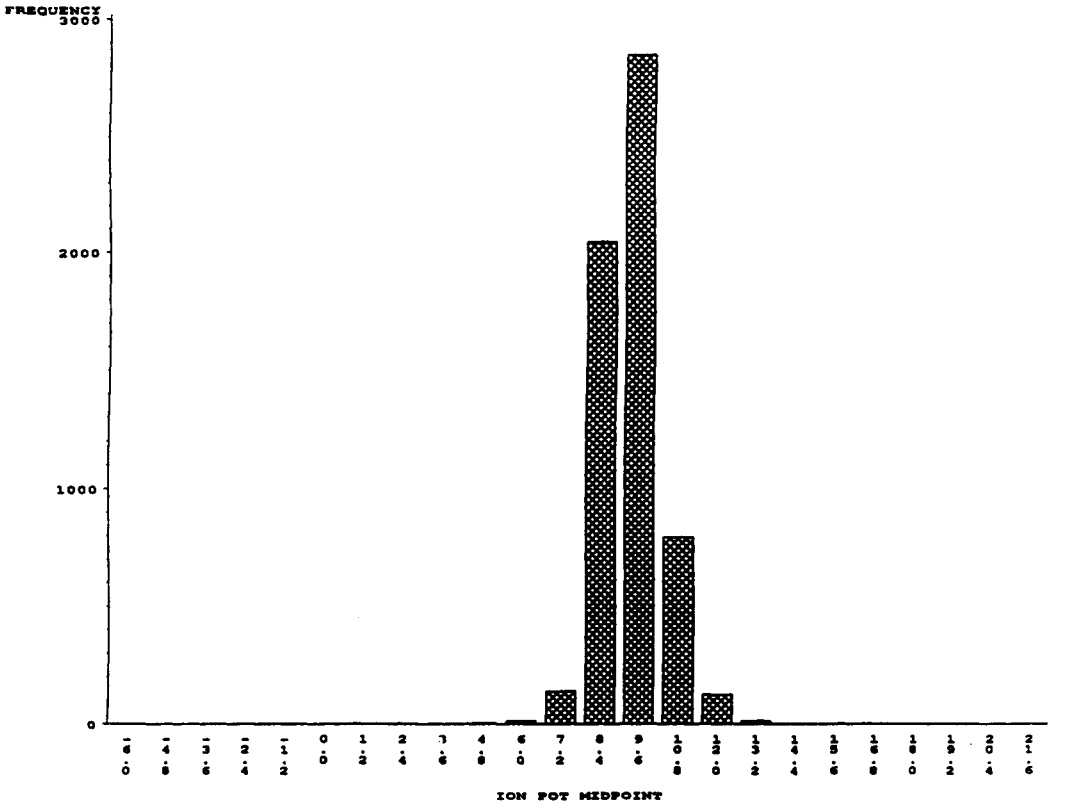


Figure 8. Ionization potential frequency distribution on experimental database.

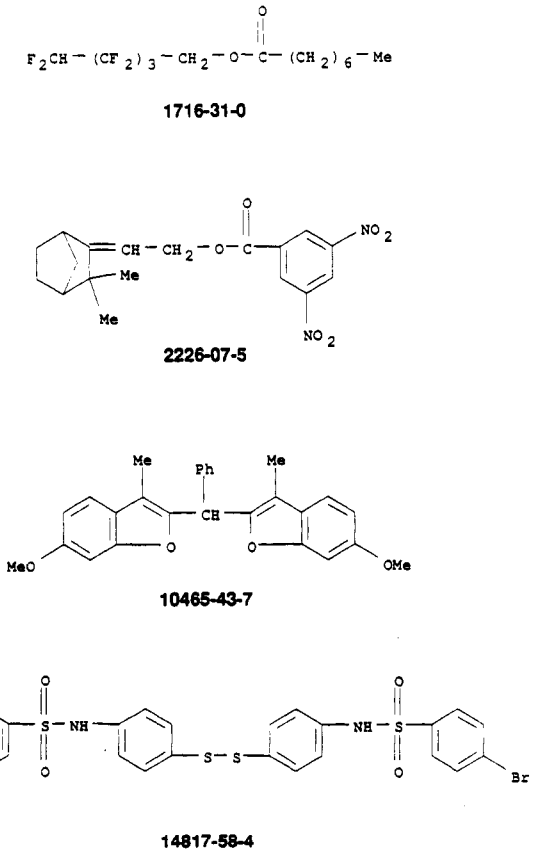


Figure 9. Query set examples.

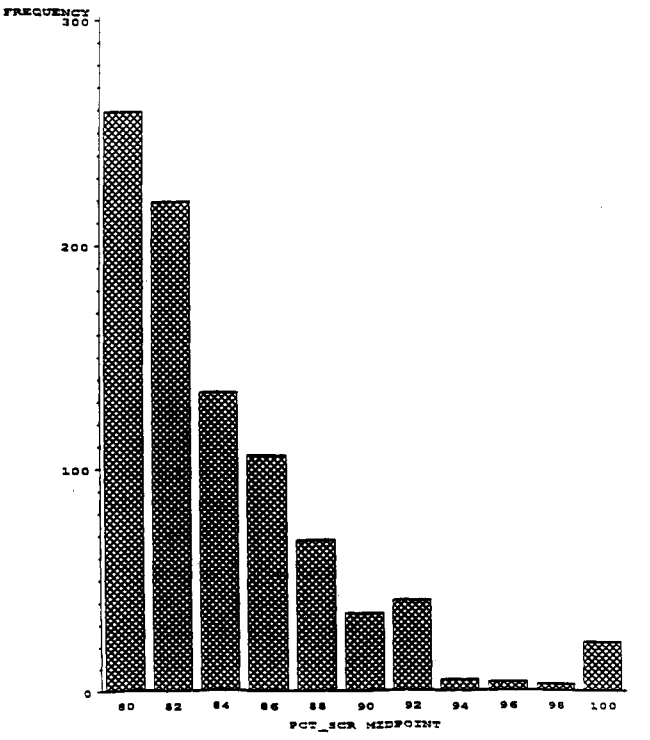


Figure 10. Percent similarity histogram (20 global molecular property queries: 896 total hits greater than 80% similarity).

system but a somewhat different environment for the hetero atoms relative to the target substance. Figure 11 illustrates another such case. The answer is the top right of the figure has a very similar 2D structure relative to the target substance, while the bottom structure in the figure which has the same similarity score has a somewhat different 2D structure. The feature values for the target and the two retrievals are shown in Table II. Although there are some feature value differences that can be attributed to the ring size difference, most of the feature values among the three substance are fairly close, especially the charge and density features. Thus, it appears that the global molecular property features might be able to detect a certain amount of chemical or isosteric similarity and, perhaps, even *bioisosteric similarity* (substances with the same biochemical or pharmacological response). The exact nature of this molecular property based similarity is not yet

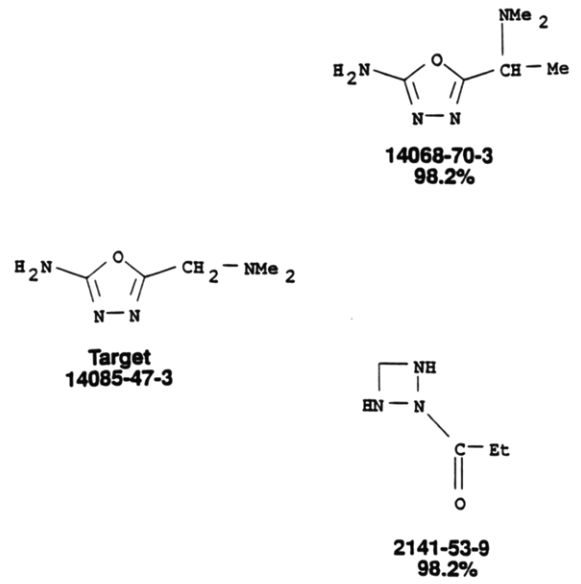


Figure 11. Illustration of "chemical similarity" in a global molecular property search.

Table II. Global Molecular Properties for Target Substance (RN 14085-47-5) and Top Two Hits

		target compd	5-ring compd	4-ring compd	database STD
1	ht_form	63.86	64.60	79.02	137.64
2	tot_eng	−1922.83	−2078.33	−1572.99	1320.16
3	ion_pot	9.37	9.42	9.60	1.00
4	dsum_t	4.11	3.05	3.68	2.43
5	lumo	0.80	0.71	0.50	1.21
6	del_h_l	−10.17	−10.13	−10.09	1.49
7	chg_mean	−5[a]	4.3[a]	−5.8[a]	0.01
8	chg_std	0.17	0.17	0.17	0.14
9	ed_mean	2.80	2.70	2.71	0.54
10	ed_std	2.04	2.00	2.06	0.20
11	ed_min	0.73	0.73	0.83	0.26
12	ed_max	6.18	6.17	6.27	0.59
13	logp	0.38	0.91	0.62	2.14
14	vol	130.34	145.82	107.68	81.71

[a] = $\times 10^{-6}$.

clear. The various charge and density features may be able to produce some *electrostatic similarity*. We are currently examining procedures for comparing the electrostatic potentials of substances. Also, we are exploring the ability of molecular property results to predict chemical and physical properties, e.g., boiling points and solubility.

Generic Triangle Metric Searching. Fuzzy-match similarity searches were also performed using a full set of 80 generic triangle features (i.e., the last six classes shown in Figure 2) and a corresponding reduced set of 29 features. Our goal in experimenting with these triangle features was to test their ability to detect shape and size similarity. Intuitively, the triangles should be better than generic atom pair distances since they contain more information. Based on the results of these searches, it appears that the generic triangle features detect a significant amount of shape and size similarity. This is illustrated in Figures 12 and 13. The wire-frame and van der Waals surface models were produced using the SYBYL molecular modeling software package.[29] In Figure 12, the substance with the highest degree of similarity has essentially the same 2D structure as the target substance and obviously the same basic shape and size. The second substance has an acyclic branch instead of a phenyl branch, but essentially the same shape and size as the target substance. The third

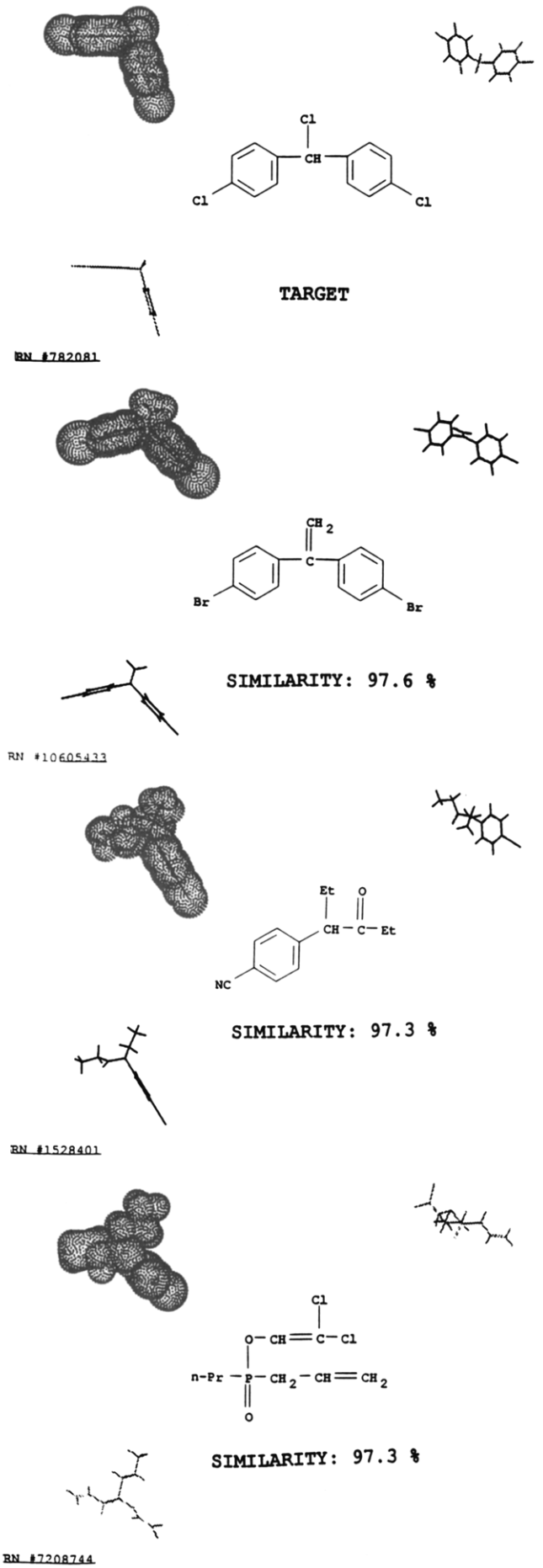


Figure 12. Illustration of "shape and size similarity" in a generic triangle feature search.

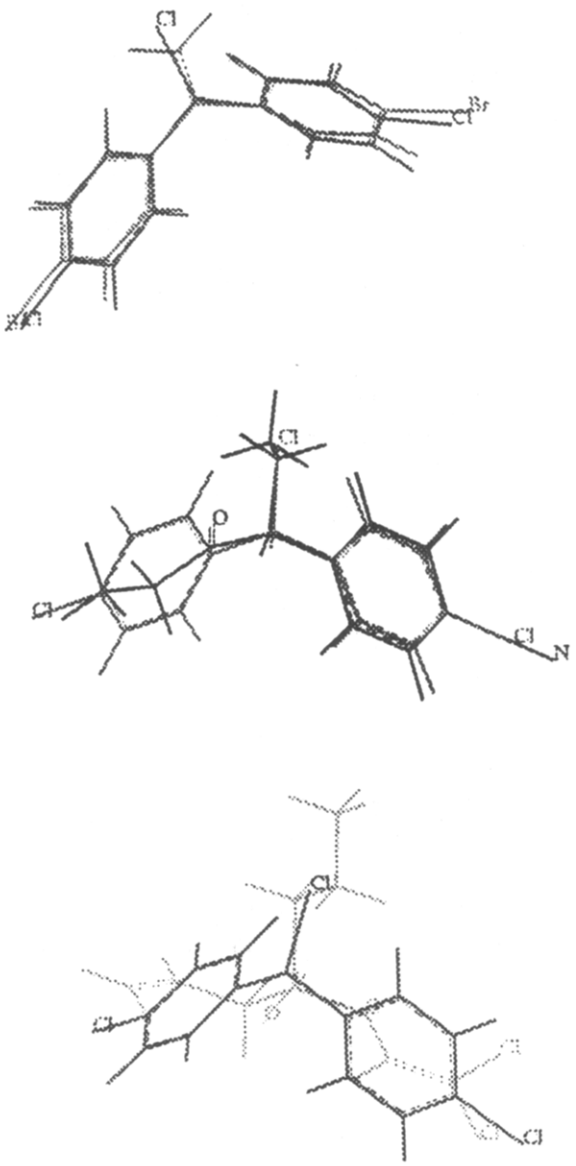


Figure 13. Superimposition of generic triangle features search top hits onto target substance.

Table III. Prediction of Volumes from Generic Triangle Search Results

query	N	target vol	predicted vol	predicted STD	mean sim, %
1	16	171.26	202.71	24.99	93.4
3	61	219.71	211.92	23.53	94.3
13	16	197.68	205.13	21.43	93.1
18	39	130.33	165.78	17.24	91.8
		Global Molecular Property Search			
13	30	197.68	206.89	29.85	91.9

substance is topologically very different from the target substance, but still has the same overall shape and size. The superimposition of the three top hits on the 3D model of the target substance is shown in Figure 13. The top substance has a root-mean-square (RMS) value of 0.13 using a 20-atom alignment; the second substance has a RMS of 0.24 using a 9-atom alignment; and the third substance has a RMS of 0.65 using a 6-atom alignment.

The shape and size of a particular conformer of a molecule are at least somewhat related to its volume. Table III presents the results of volume prediction via the generic triangle, fuzzy-match results for four, full-feature set queries. The predicted target volumes are the means of the volumes of the fuzzy-match answers with greater than a 90% similarity score. Also shown for one of the queries (no. 13) is the volume prediction via the leave one out technique on a molecular property search. The predicted volumes are very good for queries 3 and 13. For query no. 13 the molecular property and the triangle feature predictions are very similar. We plan to expand this volume prediction experiment to include a more appropriate number of target substances.

To verify that the triangle features are responsible for the shape and size similarity, we performed several searches using only nontriangle distances, i.e., Any- - -Any atom pair distances. The results were significantly different, and only occasional shape and size similarity were visualized for the atom pair distance features.

It is important to note that the generic triangle features do not reflect specific node types, i.e., the shape is independent of the kinds of atoms in the substance. It should be possible to "color" the nodes to a certain extent with the triangle atom and/or the triangle 3-slot features, and this is being investigated. These experiments illustrate the importance of having a system that allows the user to select different feature classes so that their similarity criteria can be defined.

V. SUMMARY AND CONCLUSIONS

A suite of experimental software has been developed that has proven to be very useful in modeling fragment-based similarity searching. An experimental database containing over 3000 2D, 2D/3D, 3D, and molecular property features for approximately 6000 substances has been used in conjunction with this software. A unique statistics-based scoring method has been used for the numeric features in fuzzy-match similarity searching. Perhaps the most significant characteristic of this software system is that essentially any number and combination of features can be expressed as an inclusion or fuzzy-match search. This allows for comparison of features and feature classes within a data type and between data types. Thus, the software system, to a large extent, allows the user to define their own similarity criteria for a target substance.

The fuzzy-match similarity searching of global molecular properties appears to detect chemical or isosteric similarity. Likewise, significant shape and size similarity are found from fuzzy-match searching on generic triangle 3D features. Such similarity types could be important in the course of molecular design. These new similarity types are viewed as being complementary to the more traditional 2D structural similarity. Additionally, fragment-based, fuzzy-match similarity searching is envisioned as a screening step to more detailed but time-consuming procedures, such as a shape and electrostatic potential analysis. The relatively small number of features needed for the new similarity types should make the searching of large files of substance more feasible. The input to such searching would be a 2D structure and a specification of similarity type and, optionally, individual feature and feature class identifiers.

ACKNOWLEDGMENT

We would like to thank the following members of the CAS research staff for their assistance in this project: A. H. Lipkus and G. G. Vander Stouw. Also, we would like to thank a former member of the CAS Research Department, D. H. Lillie, for his input.

REFERENCES AND NOTES

(1) Wigington, R. 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.

(2) 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.
(3) 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*, 212–218.
(4) Farmer, N. A.; O'Hara, M. P. CAS ONLINE—A New Source of Substance Information from Chemical Abstracts Service. *Database* **1980**, *3*, 10–25.
(5) 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 575–586.
(6) 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.
(7) Fisanick, W. Requirements for a System for Storage and Search of Markush Structures. In *Computer Handling of Generic Chemical Structures*; Barnard, J. M., Ed.; Gower: Aldershot, U.K., 1984; pp 106–129.
(8) Fisanick, W. The Chemical Abstracts Service Generic Chemical (Markush) Structure Storage and Retrieval Capability. 1. Basic Concepts. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 145–154.
(9) Fisanick, W. Storage and Retrieval of Generic Chemical Structure Representations. U.S. Patent 4,642,762, Feb 10, 1987.
(10) Ebe, T.; Sanderson, K. A.; Wilson, P. S. The Chemical Abstracts Service Generic Chemical (Markush) Structure Storage and Retrieval Capability. 2. The MARPAT File. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 31–36.
(11) Fisanick, W.; Cross, K. P.; Rusinko, A., III. Characteristics of Computer-Related Molecular Property Data for CAS Registry Substances. *Tetrahedron Comput. Methodol.* **1992**, in press.
(12) Cross, K. P.; Fisanick, W.; Rusinko, A., III. Comparison of Atom-by-Atom 3D Search Routines for Searching CAS Registry Substances. Presented at the 4th Chemical Congress of North America and 202nd National American Chemical Society Meeting, New York, Aug 1991.
(13) Cross, K. P.; Fisanick, W.; Rusinko, A., III. Searching Rigid 3D Structures as Templates for Chemical Syntheses. Presented at the 24th Central Regional Meeting of the American Chemical Society, Cincinnat, OH, May 1992.
(14) (a) Rusinko, A., III; Skell, J. M.; Balducci, R.; Pearlman, R. S. *CONCORD User's Manual*; Tripos Associates: St. Louis, MO. (b) Pearlman, R. S. Rapid Generation of High Quality Approximate 3-D Molecular Structures. *Chem. Des. Auto. News* **1987**, *2* (1), 1, 5–6. (c) Rusinko, A., II. Tools for Computer-Assisted Drug Design. Ph.D. Thesis, The University of Texas, Austin, TX, 1988.
(15) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902.
(16) Willett, P. *Similarity and Clustering in Chemical Information Systems*; Research Studies Press: Letchworth, Hertfordshire, England, 1987.
(17) Johnson, M. A.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; John Wiley and Sons: New York, 1989.
(18) Pepperrell, C. A.; Willett, P. Techniques for the calculation of three-dimensional structural similarity using inter-atomic distances. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 455–474.
(19) Pepperrell, C. A.; Poirrette, A. R.; Willett, P.; Taylor, R. Development Of An Atom Mapping Procedure for Similarity Searching In Databases of Three-Dimensional Chemical Structures. *Pestic. Sci.* **1991**, *33* (1), 97–111.
(20) The SAS System is available from the SAS Institute, Inc., SAS Circle, P.O. Box 8000, Cary, NC 27512-8000.
(21) The MOLCONN2 software is available from Dr. Lowell H. Hall, Hall Associates Consulting, 2 Davis Street, Quincy, MA 02170.
(22) Kier, L. B. A Shape Index from Molecular Graphs. *Quant. Struct.–Act. Relat.* **1985**, *4* (3), 109–116.
(23) Shannon, C. E.; Weaver, W. *The Mathematical Theory of Communication*; University of Illinois Press: Urbana, IL, 1949.
(24) Kier, L. B. An Index of Molecular Flexibility from Kappa Shape Attributes. *Quant. Struct.–Act. Relat.* **1989**, *8*, 218–221.
(25) 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.
(26) Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. Atomic Physicochemical Parameters for Three-Dimensional Structure Directed Quantitative Structure–Activity Relationships. 4. Additional Parameters for Hydrophobic and Dispersive Interactions and Their Application for an Automated Superposition of Certain Naturally Occurring Nucleoside Antibiotics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163–172.
(27) Pearlman, R. S. Molecular Surfaces and Volumes and Their Use in Structure/Activity Relationships. In *Physical Chemical Properties of Drugs*; Yalkowsky, S. H., Sinkula, A. A., Valvani, S. C. Eds.; Marcel Dekker: New York, 1980.
(28) The use of cluster analysis techniques to produce a set which is independent of ordering has been suggested by one of the referees of this paper.
(29) SYBYL is a molecular modeling software suite available from Tripos Associates, Inc., 1699 S. Hanley Road, Suite 303, St. Louis, MO 63144-2913.