Similarity Searching on CAS Registry Substances. 2. 2D Structural Similarity[†]

William Fisanick,* Alan H. Lipkus, and Andrew Rusinko III

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

Received July 21, 1993[⊗]

Chemical Abstracts Service (CAS) is exploring approaches for similarity ("fuzzy-match") searching on CAS Registry substances. Experimental software is being developed to identify, analyze, and perform similarity searches on various characteristics of an integrated set of 2D, 3D, and molecular property data for samples of Registry substances. Earlier results have indicated that searching on global molecular property features such as ionization potentials and van der Waals' volumes appears to detect "chemical" (isosteric) similarity and that searching on generic atom triangle geometric features provides a significant amount of shape and size similarity. More recently, we have been exploring possibilities for 2D global and local structural similarity on Science and Technology Network (STN) structure files. One possible approach involves one or more fragment-based searches using the existing STN 2D substructure screen fragments, optionally followed by existing connectivity-based (atom-by-atom) search on generic structure representations of candidates obtained in the fragment-based screening step. Fragment-based searches using various screen classes such as augmented atoms and bond sequences provide for different views of 2D structural similarity. Connectivity-based searching of generic 2D structures allows for a considerable amount of flexibility in a user's definition of similarity. This paper will discuss recent results of a comparison of the effectiveness of the various STN screen classes in fragment-based similarity searching using the Tanimoto coefficient and will illustrate the STN capabilities for connectivity-based similarity searching on an answer set.

INTRODUCTION

Chemical similarity searching (sometimes referred to as "fuzzy-match" searching) is rapidly becoming a powerful source of useful and interesting information to chemists. It has been particularly useful in the areas of computer-aided synthetic planning[1,2] and molecular design.[3] Recent research at Chemical Abstracts Service (CAS) has led to the development of an experimental system which searches integrated 2D (topology-based), 3D (atomic coordinate and geometric features), and molecular property (such as the ionization potential, van der Waals' volume, etc.) data generated from the CAS Registry File. Previous results from testing of this system have indicated that searching on global molecular property features appears to detect chemical "isosteric" similarity[4] (perhaps even "bio-isoteric" similarity[5]), and that searching on generic atom triangle geometric features provides a significant amount of shape and size similarity.[6,7] More recently, we have observed that generic atom "tetrangles", which consist of a bond vector pair and the interatomic distances between the base atoms and between the tip atoms, can also detect shape and size similarity. However, until recently, research at CAS has not been focused on the problems associated with the development of a 2D similarity search system.

The concept of similarity is not one that is well-defined since there may be considerable variance among users in their perspectives and in their objectives in using such a capability. Thus, the development of a chemical similarity search system requires of a balance between flexibility (allowing the user controls for many options) and practicality (what can be reasonably accomplished on a system). A possible scenario for the development of a 2D similarity search system utilizes one or more fragment-based searches on existing Science and Technology Network (STN) 2D substructure screen fragments to rank structures versus a target structure. This primary search could be optionally followed by a generic or generic/real structure search that identifies substances more clearly related "connectivity-wise" to the target structure on the basis of the entire structure (global) or a portion of it (local). Fragment-based search using various STN screen classes such as augmented atoms or bond sequences permits different views of 2D structural similarity, while connectivity-based searching of generic 2D structures on a resultant answer set allows for a considerable amount of flexibility in a user's definition of similarity. Furthermore, different fragment weighting schemes could be employed which also allow the user finer control of the results from a similarity search. Therefore, substances highly similar to the target compound on a global or local basis can be identified quickly using Boolean operations on bit-mapped screens and precisely via a secondary generic structure atom-by-atom search.

The first phase of our similarity search employs a fragment-based similarity scheme analogous to those used by many others.[8,9] Our similarity will be different in that STN structure search screens will be used as the search fragments or features. Currently on STN there are 2127 different structure search screens defined for use in the screening step prior to performing a 2D structure search. Use of search screens for similarity measures has been reviewed by Willett,[10] and discussions of many similarity search systems can be found in refs 10–12. On the other hand, the second-phase search is conceptually related to the reduced graph representation search of Takahashi et al.[13] Also, Hagadone has reported on a system that uses a screening level and a refined search level which utilizes a maximum common substructure search.[14] This report describes the work done on one aspect of the experimental integrated system, namely, our initial efforts to define a 2D structure similarity system for STN structure files.

[†] Presented in part at the Third International Conference on Chemical Structures held in Noordwijkerhout, The Netherlands, June 6–10, 1993.

[⊗] Abstract published in *Advance ACS Abstracts*, January 15, 1994.

0095-2338/94/1634-0130$04.50/0

STN 2D STRUCTURE REPRESENTATION AND SIMILARITY

In search and retrieval on STN structure files, chemical structures are represented by a set of predetermined substructure screens which encode key features of a structure and a connection table representation which fully encodes the structure. In a substructure search, the screens are used in a preliminary or screening step to obtain a set of candidate substances which are then processed in a precision-refining, atom-by-atom search on their connection tables. Both types of representations have potential for use in similarity searching and are described below.

Substructure Screen Fragments. The STN substructure screen set encodes over 5400 unique structural fragments in over 2100 screens.[15] Screen number "sharing" is used to allow several fragments to be represented by a single screen. The fragments are "ORed" together for the screen; i.e., if any one of them is found in a structure the screen is "set". Typically, screen sharing is used to collect together fragments with related atoms, for example, "Br, Cl, F, or I"; "O or S"; or uncommon hetero (As, B, P, Se, Si, Te). Another typical screen sharing is for fragment sets differing by the combinations of single, double, and normalized bonds.

The focus of the STN screen set is on "rarer" structural fragments or fragment sets since its purpose is to effect as high a screenout in substructure searching as possible. For example, on a sample file of 30 K substances (see below), 50% of the screens have a frequency of occurrence of less than 2.1%; for 75% of the screens, it is less than 6.4%.

There are 12 screen classes in the STN screen set: augmented atoms (AA), hydrogen augmented atoms (HA), twin augmented atoms (TW), atom sequences (AS), bond sequences (BS), connectivity sequences (CS), ring count (RC), type of ring (TR), atom count (AC), degree of connectivity (DC), element composition (EC), and graph modifier (GM). The screen classes are fully described in ref 15.

A priori, the STN screen set would seem to possess several properties that should allow the effective fragment-based similarity searching. The screen set was designed for an initial environment that involved manual encoding of screen numbers in a Boolean expression. Because of this environment, a fairly large number of precise screens were included in the set, such as augmented atoms with both bond type and values—the user would handle bonding vagueness in a query by ORing together several precise screens. Also, the precise CS's and BS's could be handled in the same manner. Such precise screens are typically not generated automatically from a substructure query, because of the potential for recall failure. However, since the target for a similarity search is a complete structure, these precise screens would be valid, and, indeed, should improve the effectiveness of the search results. Also, the STN screen set has a fairly large number of screen classes which should tend to give some balance to the set. Clearly, the dominant classes are the AA and AS classes. However, there may be features in certain queries that are not handled adequately by the AA's and AS's (see below).

There would also seem to be a potential to obtain different "similarity views" by using the screen classes individually or in combinations of a few classes. For example, the AA and AS classes should provide an overall or "general" similarity; the CS and DC classes, a "connectivity" similarity; the HA and TW classes, a "substitution" similarity; the RC and TR classes, a "ring" similarity; the AC class, a "size" similarity; and the EC and element-GM classes, an "atom" similarity.

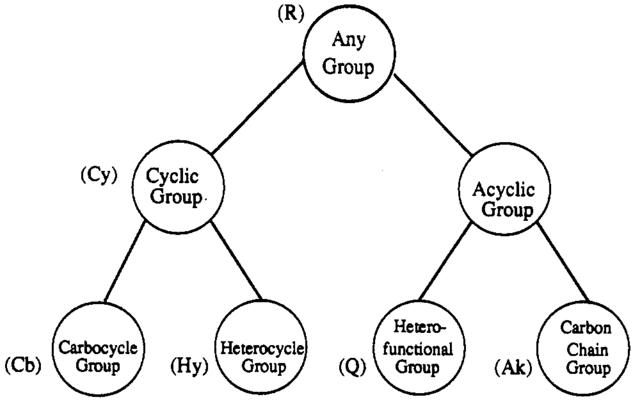


Figure 1. Generic group hierarchy used in STN structure searching.

One aspect of the STN screen set which might reduce its effectiveness as a feature set for similarity searching is sharing of a screen number by a set of dissimilar fragments. Fortunately, however, there are only a small number of such screens. The redundancy due to screen fragment degeneracy might also have a negative effect. Also, the lack of multiple fragment counts for certain screen classes might tend to reduce the effectiveness.

An evaluation of the effectiveness of the STN screen set with respect to 2D structural similarity searching is given in a later section.

Structure Connection Tables. The structure connection tables used in STN search and retrieval can be viewed as a specific–generic continuum with respect to representation. At the generic end of the continuum, a hierarchical set of generic groups is used. This set of groups was originally developed for use in Markush structure handling.[16–19] This hierarchy is illustrated in Figure 1. The generic groups are specified in the circles. The symbols within parentheses to the left of a group are valid STN structure node symbols.

The most generic group is the "any" group (R), i.e., any group of atoms. This group is divided into a "cyclic" group (Cy) and an "acyclic" group. A cyclic group may be a "carbocycle" group (Cb) or a "heterocycle" group (Hy). An acyclic group may be a "heterofunctional" group (Q) which is simply a heteroatom and any attached hydrogens or a "carbon chain" group (Ak) which can be defined an acyclic group of carbon atoms delimited by a cyclic group, Q or "null". The hierarchy has several properties such as the ability to cover all portions of chemical structure and, because of the fairly simplistic definitions, the potential for use in structure query-framing by even novice users.

The various generic groups in the hierarchy can be used as "windows" to view sets of similar real-atom fragments. This is illustrated in Figure 2 for the heterocycle group. Without any qualification, heterocycle will view all heterocyclic ring systems, such as those illustrated on the right of the Figure 2, from a system as small as ethylene oxide to one as large as a porphyrin. The view of a generic group such as heterocycle is probably too broad to be useful, but a more narrowly defined group such heterofunctional can be appropriate. For a group like heterocycle, qualifying the group is probably necessary; i.e., one partly closes the window. Currently, on STN such qualifications can be made via a set of generic group categories and attributes. For example, one can specify that the heterocycle be restricted to a saturated monocycle with exactly two carbons and one oxygen. This would probably narrow the heterocycle to just one ring system, ethylene oxide. Similarly, specifying an unsaturated monocycle with three to four carbons and two to three nitrogens views a family of ring

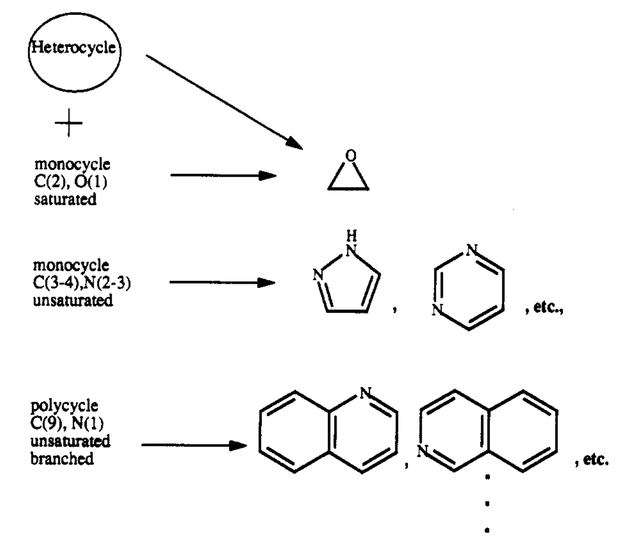


Figure 2. Illustration of the similarity view of real-atom structural fragments via the heterocycle genetic group.

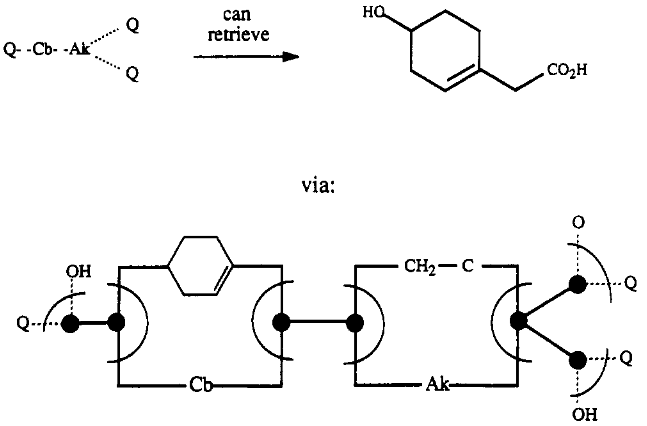


Figure 3. Illustration of the matching of generic group query structure with a real-atom structure on STN specific structure files.

systems consisting of pyrazole, pyrimidine, etc.; specifying an unsaturated, branched polycycle with exactly nine carbons and one nitrogen gives a family which includes the quinolines shown in the last row of Figure 2.

Thus, with the current set of generic group categories and attributes, there is a significant amount of control in the view of similar real-atom fragments via a generic group node.

Currently on STN, generic structure queries such as the one shown in the upper left-hand side of Figure 3 can be framed. This query can retrieve the hydroxycyclohexeneacetic acid shown in the upper right-hand side of Figure 3. There is a mapping of the Q on the Cb to hydroxy, the Cb to cyclohexene, the Ak to the carbons in the acetic acid, and the two Q's on the Ak to the oxygens in the acetic acid.

In structure searches on STN specific structure files such as the CAS Registry File, if a structure query contains a generic group node, a "composite" structure is generated and matched against at atom-by-atom search time.[17] The composite structure for the acetic acid derivative is shown on the bottom of Figure 3. For each real-atom fragment, an alternative generic group node is generated. The "arcs" in Figure 3 indicate alternatives; the solid dots are "null" connectors. Thus, the generic structure query matches the hydroxycyclohexeneacetic acid file substance because a path can be traced from Q to Cb to Ak to the two Q's.

Generic structures such as the one in Figure 3 are useful in collecting together a set of "connected" or "partially connected" similar real-atom structures via the "windowing" effort of the generic groups. However, such a generic structure query currently will not execute against the full CAS Registry File on STN since only real-atom screens are used in searching and none would be available for this query. The query can be executed, however, in a "subset" search of up to 100K answers derived in a preceding search.

Thus, if a traditional fragment-based similarity search were available to generate a set of candidate answers, this type of connectivity-based similarity search could be used to provide for greater precision in similarity definition.

A description of STN connectivity-based searching with respect to 2D structural similarity is given in a later section.

SEARCH APPROACH

As mentioned in the Introduction section, a desirable 2D structural similarity capability is one that that can provide for a range of similarity views in an environment amenable to both novice and expert users. We have been investigating the feasibility of a STN capability we believe will meet these objectives to a large extent. The capability would utilize the STN substructure screen set or subsets of it for fragment-based similarity using the Tanimoto similarity coefficient. This coefficient can be defined as

$$T = \frac{C}{(Q + \mathrm{FS}) - C}$$

where T is the coefficient, C is the number of screens in common between the query and file substance, Q is the number of query screens, and FS is the number of file substance screens. Also, as part of this capability, secondary fragment-based and connectivity-based similarity searches could be executed as subset searches on answer sets to provide different similarity views and more precise similarity definitions on a global or local basis. The anticipated processing flow for 2D structural similarity on STN specific structure files is shown in Figure 4.

A typical novice scenario would be one in which the user retrieves the structure of a target substance via a Registry Number RECALL, requests a fragment-based similarity search for the target structure using the default, "general" screen set, and browses the answer set ranked in the order of the Tanimoto similarity coefficient. As with substructure search, typically a SAMPLE search would be executed first to determine a reasonable similarity coefficient threshold. The FULL search would then be executed using this threshold, and ranked answers would be accumulated up to the maximum answer file size (100K substances) within the threshold limits. If the target structure is recalled from the database, then a 2D structural similarity search is simpler than a substructure search since the query structure need not be built; it would be comparable to a substructure search in complexity if the target structure had to be built.

One could obtain different global similarity views for a target structure by executing a subset similarity search on the initial answer set using the screen classes corresponding to the view, e.g., the BS class for a "bonding" view. The results of this secondary class searching would be a reranking of the initial answer set from a "bonding" viewpoint. Alternatively, individual class or class combination searches may be appropriate against the full file. Also, answer sets could be reranked on a local basis by specifying that portions of the target structure receive a higher weight in computing the coefficient and then executing a secondary similarity search on the initial answer set with the weighted query. Thus, file

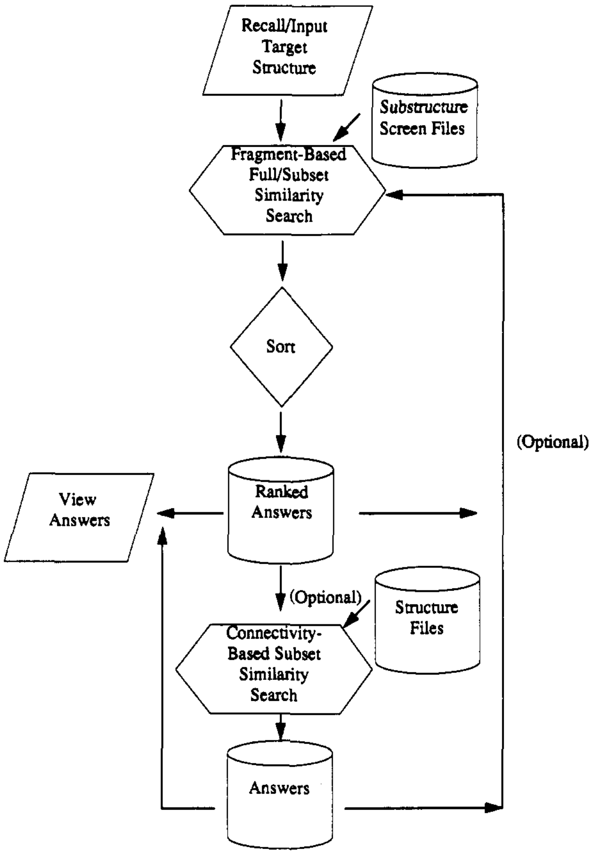

Recall/Input Target Structure
Substructure Screen Files
Fragment-Based Full/Subset Similarity Search
Sort
(Optional)
View Answers
Ranked Answers
(Optional)
Structure Files
Connectivity-Based Subset Similarity Search
Answers

Figure 4. Anticipated Processing Flow for 2D Structural Similarity on STN Specific Structure Files.

structures which have the specified portions will tend to be in the initial part of the ordered answer set.

An optional, connectivity-based similarity search using a generic query structure would be available for a more precise similarity definition on the initial or subsequent fragment-based search results. A hierarchical generic group structure, or any portion of it, would be defined by the user or, possibly, selected from several types of generic structures generated automatically from the target structure. As mentioned above, such generic structures can currently be built on STN. The results of such a search would be a smaller number of more precise answers which could be reranked by executing on the answer subset a fragment-based similarity search or, possibly, via a new connectivity-similarity measure we are exploring which is described in a later section.

The key question with respect to the feasibility of this search approach is the effectiveness of the STN substructure screen set for 2D structural similarity via the Tanimoto coefficient. Can a general, default set of screens be identified that is effective with respect to precision and recall; how effective are the individual screen classes in connoting different similarity views, etc.? If these screens prove to be effective, then it should be possible to implement a 2D structural similarity search capability on STN fairly quickly since the screen fragment set is already available.

EXPERIMENTAL SEARCH AND RETRIEVAL SYSTEM

The components of an experimental search and retrieval system that was used in assessing the feasibility of the above approach are illustrated in Figure 5. These are loosely coupled components which are interfaced via Registry Number answer files that are transmitted using CAS's local area network

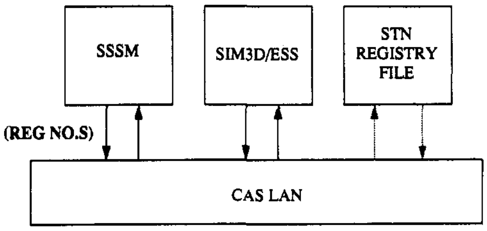


Figure 5. Components on an experimental search and retrieval system used for similarity searching.

(LAN). There is a pilot system called the Similarity and 3D Searching Experimental Search System (SIM3D/ESS). This pilot system is currently being used by users external to CAS as part of an experimental search service for 3D substructure searching on a collection of databases containing over a half-million 3D structures.[6] Currently, the SIM3D/ESS is used as a data source and for display purposes for similarity searching experiments. As the similarity search capabilities mature, the intention is to include the more promising ones in the SIM3D/ESS.

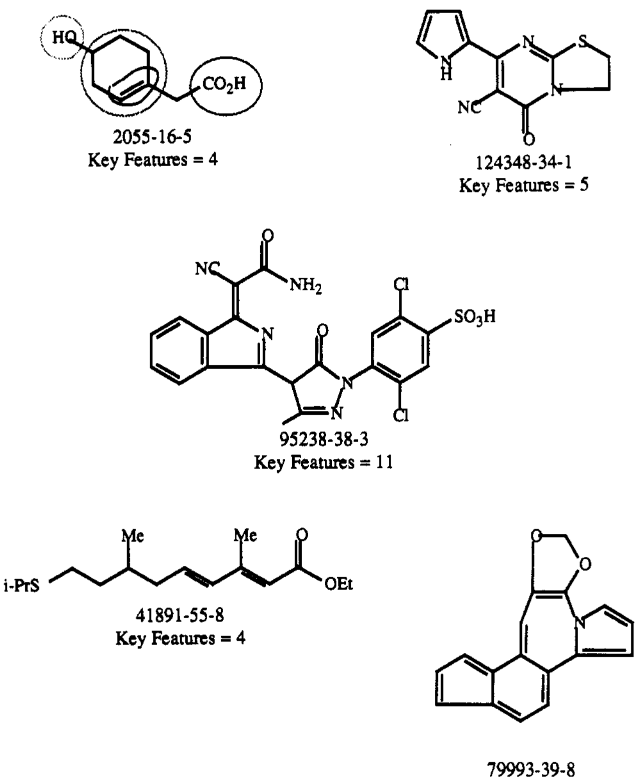
The fragment-based similarity searching is accomplished using a collection of software and files which we call the Substance Similarity Search Modeler (SSSM).[4,7] The SSSM is also used to develop similarity search techniques and to study relationships among 2D, 3D, and molecular property features.

The CAS Registry File on STN is used for connectivity-based similarity searching. Answer files from the fragment-based searching on the modeler are transferred as an answer set on the Registry File and are used in subsequent subset searches for generic group structure queries. A script file is used to automatically log on to the Registry File, search for Registry Numbers from the modeler search, and to save the resultant Registry File answer set. The connectivity-based searches are then executed against this answer set in a subsequent session.

The modeler is written in SAS (Statistical Analysis System[20]) language (UNIX environment) and, thus, all of the SAS statistical procedures are available to it. A variety of similarity techniques are supported, including two statistical techniques that we have developed.[4,7] These statistical techniques will automatically generate sets of numeric search ranges for each feature specified in the query based on the standard deviation from the mean of a feature for the substances on the database or some suitable sample of these substances. These statistical methods tend to work best for "dense" feature sets, i.e., when the features have values of >1 for most of the database substances.

The Tanimoto and Euclidean distance methods are also available. We tend to favor the Tanimoto method for "sparse" feature sets, such as the STN substructure screen set. In a sample file of 30K substances with STN screens the average number of screens set was 135 out of a possible 2127 screens, i.e., the STN substructure screens are a sparse set of features. The Euclidean distance method is used mostly for dense feature sets.

Individual feature, feature class, and database frequency weighting are supported in the modeler. However, the most significant feature of the modeler is that any combination of features can be used for the query, from a single feature up to the number available for each substance on the database. This allows us to readily study the performance of various subsets of the total set of features with respect to similarity search. This total set may be an integrated set of 2D, 3D, and molecular property features.



2055-16-5
Key Features = 4

124348-34-1
Key Features = 5

95238-38-3
Key Features = 11

41891-55-8
Key Features = 4

79993-39-8
Key Features = 4

Figure 6. "Pilot-5" set of similarity search queries.

Two databases were used in experimental similarity searching on the STN substructure screens: (1) a 60 000 substance file (60K file), which was derived by a systematic sampling of the 4.5 million Registry substances that had 3D coordinates when the file was generated,[6] and (2) a 30 000 substance file (30K file) which contains every other substance on the 60K File. Both these files contain 2228 features per substance, including a zero for the absence of a feature. In addition to the 2127 STN substructure screens, there are 32 global molecular properties such as ionization potential and log P, 27 generic 2D features such the count values of the bond types and values, and 42 generic augmented atom count values such as the number of 4-atom augmented atoms (i.e., A A A A). These non-STN features were used in cross checking of search results and as targets for feature prediction which is described in the next section.

A set of 20 queries was used in the testing. This set is a systematic sample of the 30K file. Most of the searching was conducted on a subset of five queries which we call the "pilot-5" set. The pilot-5 set is illustrated in Figure 6. The three structures at the top part of Figure 6 represent "small", "medium", and "large" structures with respect to size and functionality. The two structures on the bottom of Figure 6 include a large all chain structure and a large all ring structure. We have identified what we consider to be the key features in each of the structures. The key features are illustrated for the structure of the upper-left of Figure 6. They are the hydroxy group, the ring, the ring unsaturation, and the carboxy group. These features were used in a manual determination of retrieval precision which is described in the next section.

The various STN screen classes and class combinations that we have used, thus far, as feature sets in similarity are illustrated in Figure 7. With the Tanimoto method, it is necessary to generate and use in searching the "Tanimoto sum vector" for a feature set being searched upon. This sum vector has the FS value in the Tanimoto equation (see above) for the database substance.

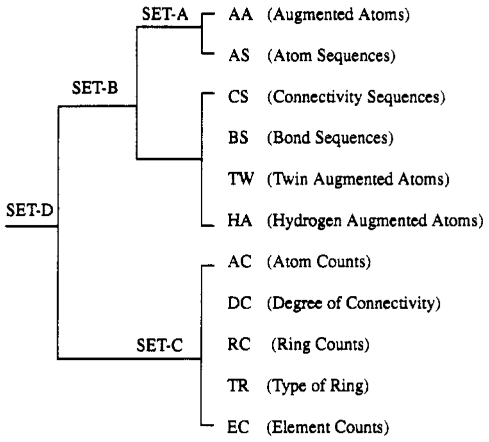


Figure 7. STN screen classes and class combinations used as feature sets in similarity searching.

- Database Frequency Weighting

SET- D:	
Freq. Range	Weight
0.0 - 0.4	10
0.4 - 1.0	8
1.0 - 2.2	6
2.2 - 4.4	4
4.4 - 10.0	2
10.0-100.0	1

- Class Weighting (SET-D)

Class	AA	AS	BS	CS
Weight	1	1	5	3

- Individual Feature Weighting

Feature	-C(=O)O-
Weight	5

Figure 8. Illustration of database frequency weighting, class weighting, and individual feature weighting.

A total of 15 feature sets have, thus far, been identified and used in searching. Each individual screen class is a feature set, the exception being the GM class which is not pertinent to our sample databases since they contain only single component substances. The combination of the AA's and AS's are grouped into a feature set called SET-A which contains over 1300 of the 2127 STN screens. SET-A and the remaining augmented atoms and linear sequences, i.e., TW, HA, CS, BS, are grouped together into SET-B (approximately 1900 screens). The generic screen classes (AC, DC, RC, TR, EC) are grouped together in the SET-C feature set. The combination of SET-B and SET-C gives SET-D, i.e., the "full" set.

We are currently exploring ways to select a subset of screens from a screen class (e.g., AA) or class combination for use as a feature set. We plan to test this new feature set to see if there is a performance improvement relative to feature sets which use all the screens from a screen class or class combination.

The weighting schemes that we have been experimenting with are illustrated in Figure 8. The table used for database frequency weighting with the SET-D (full) screen feature set is shown at the top part of the figure. This table was developed by taking the frequency of occurrence of the screens on the 30K file and dividing the total number of screens into six bins of approximately 350 screens each. Thus, the first 350 screens have a frequency of occurrence of 0.0–0.4%; the second 350 screens have a frequency of occurrence of 0.4–1.0%, etc. The weights assigned to each bin are shown in the right-hand column of the table. The "rarer" bins receive a greater weight; i.e., the assumption here is that the rarer screens are more important. A weighted Tanimoto sum vector is created for the database substance using the table, i.e., larger FS values are used. The weighting table can also be used for only the query features, i.e., with the unweighted Tanimoto sum vector. This will tend to force file structures with rarer features to the top of the ranked answer set.

In a class weighting scheme, the weights for the individual screen classes in the SET-D feature set are specified by the user. All the individual screens within a screen class are automatically assigned the class weight (the default weight is 1). In searching, a class weighted SET-D sum vector can be generated automatically at search time or the unweighted SET-D sum vector can be used. Class weighting can be used to provide different similarity views. For example, in the illustration in Figure 8, a combination bonding and connectivity view is being requested with the bonding view being more heavily weighted (i.e., 5 vs 3).

The database frequency and class weighting are global weighting schemes in that they apply to the entire structure. Individual feature weighting is a local scheme in that it allows one to focus on portions of a structure rather than the total structure. Currently, individual feature weighting in the modeler is accomplished manually by specifying the weight for the appropriate query screen features. However, this could be automated by specifying appropriate nodes in the query structure. We currently use only an unweighted Tanimoto sum vector with individual screen weighting. Thus, the main use for individual feature weighting is for the reranking of answer sets in which the selected features, for the most part, would be present in the first answers of the ranked set. For example, in the illustration in Figure 8, the "ester" feature is given a weight of 5 relative to the other features which have a weight of 1. This type of weighting should tend to result in top ranked structures with an ester group.

EXPERIMENTS AND RESULTS

Our objectives in investigating the feasibility of global and local similarity searching using the STN substructure screens and generic structure searching capabilities were (1) to determine the best screen class combination for use as the overall or default feature set by examining the precision and recall for various class combinations, (2) to explore the potential usefulness for different similarity views via weighting schemes or secondary similarity searching using individual screen class feature sets, and (3) to explore the potential usefulness for more precise similarity views via generic structure searching.

Fragment-Based Precision Assessment. The mean Tanimoto score for the top ranked retrievals from similarity searches can serve as a rough indication of the precision of answers. Figure 9 gives the mean Tanimoto scores (%) for the top 20 and 349 ranked retrievals for the 20 query set and using the SET-A (AA and AS) screen class combination. Fragment-based (screen) searches were executed against the 30K file. Most of the queries have mean scores in the 50–65% range for the top 20 retrievals. As expected, the mean scores for the top 20 are higher than those for the top 349. The plot pattern of the top 20 and top 349 retrievals are the same with a few exceptions.

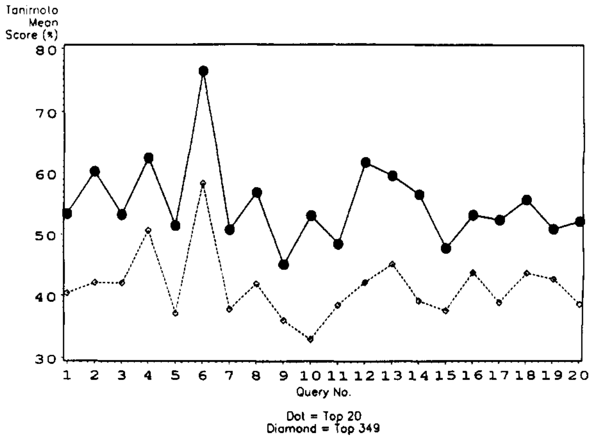


Figure 9. Mean Tanimoto score (%) for the top 20 and top 349 ranked retrievals for the 20 query set and using the SET-A feature set.

An important question to be answered is what is the impact of the scale-up to larger files? Will there be a significant increase in the scores of the top ranked answers; i.e., will the top ranked answers be significantly more precise? To gain some insight on this issue we ran the pilot-5 query set using the SET-D (full) feature set against both the 30K file and the 60K file. The mean Tanimoto percent scores increased for all 5 queries for both the top 20 and 349 ranked retrievals in going from the 30K file to the 60K file. For the top 20 retrievals the increases ranged from 2.5 to 4.3% with a mean value of 3.1%; for the top 349, the range was from 2.0 to 4.5% with a mean of 3.2%. The increase in mean scores is expected, but more experimentation is need to obtain a better handle on the magnitude; i.e., will there be an approximately 3% score increase when the file size is doubled? It should be noted that on the CAS Registry File there may be small clusters of similar substances within a set of contiguous Registry Numbers due to the indexing of a journal or patent article with similar substances. The systematic sampling used in creating the sample files should typically pick only one substance from such a cluster.

One technique we have used to obtain an estimation of the relevancy of the top ranked answers in retrieval sets is the "prediction" of feature values for the query structures. Five non-STN, generic 2D and augmented atom features were predicted for the pilot-5 query set using SET-A, SET-D, and frequency weighted SET-D feature sets. More specificially, the features predicted were the carbon count, the single bond count, and the counts of four-atom augmented atoms (A A A A), three-atom augmented atoms (A A A), and two-atom augmented atoms (A A). It should be noted that, unlike STN augmented atoms, these generic augmented atoms are "natural" in that they are not the result of a decomposition; i.e., they are graph invariants. Only the STN screen feature sets were used in the searching; i.e., the features to be predicted were not included. The predicted query feature value was taken as the mean value of that feature for the top 20 similar substances. The predicted values were compared to the actual values for the query structure and Δ (difference) values were computed. The smaller the Δ value, presumably the more relevant the answer set, although Δ values of zero or nearly zero may be uninteresting in that there may not be enough variation relative to the query structure.

Figure 10 is a plot of the Δ values (actual − predicted) against the pilot-5 query set for the four-atom augmented atom feature and using SET-A (AA and AS), SET-D (full), and SET-D frequency weighted feature sets. The full set

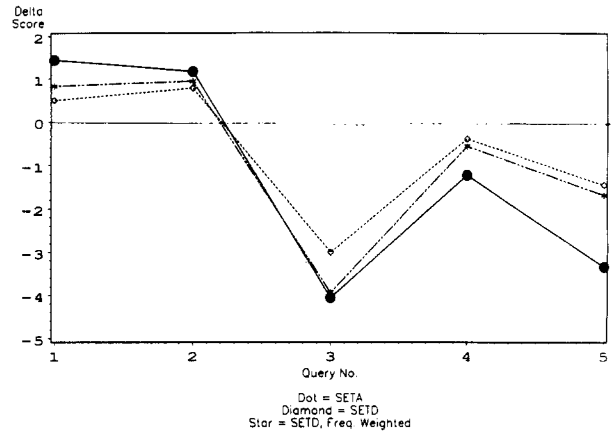


Figure 10. Top 20 Δ value (actual − predicted) for four-atom augmented atom counts of the pilot-5 query set using SET-A, SET-D, and SET-D frequency weighted feature sets.

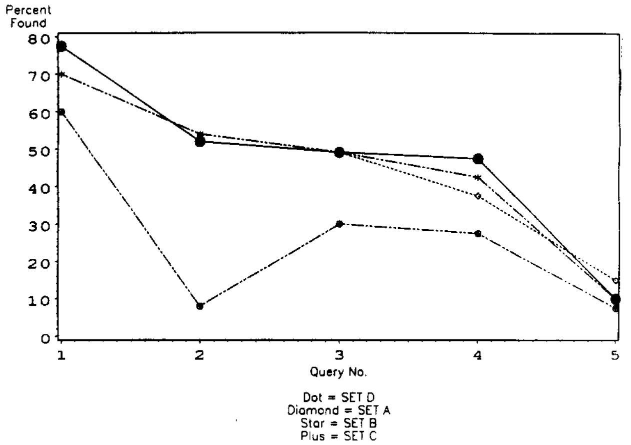


Figure 11. Mean percent of key features found in top 20 retrievals for the pilot-5 query set using SET-A, SET-B, SET-C, and SET-D feature sets.

(SET-D) seems to result in the most relevant answer set on the basis of smaller Δ values. Similarly, SET-D had the smallest Δ values in the prediction of the four other features.

We also performed a manual relevancy assessment of the top 20 retrievals for the pilot-5 query set using the SET-A, SET-B, SET-C, and SET-D feature sets. The answers were examined and the number of key features or their surrogates relative to the query counted. Also counted was the total number of ring, functional group, etc., features present in the retrieved structures. In addition, a connectivity score which reflects how well key feature fragments are connected relative to the query structure was assigned to each retrieved structure. The score values ranged from 0 (no connectivity among the fragments) to 3 (same connectivity among the fragments as in the target structure).

Figure 11 is a plot of the mean percent of key features found in the top 20 retrievals for the pilot-5 query set using the SET-A, SET-B, SET-C, and SET-D feature sets. SET-A, SET-B, and SET-D gave similar results with SET-D being slightly better overall. The retrievals from generic SET-C searches, as expected, had significantly fewer key features than the other three sets. The ratios of the total features to the key features found were very similar for SET-A, SET-B, and SET-D; i.e., there was little differentiation, but the ratios for SET-C were significantly higher.

SET-D had the greatest amount of connectivity among the key features (or their surrogates). This is illustrated in Figure 12. SET-C had the least amount of connectivity.

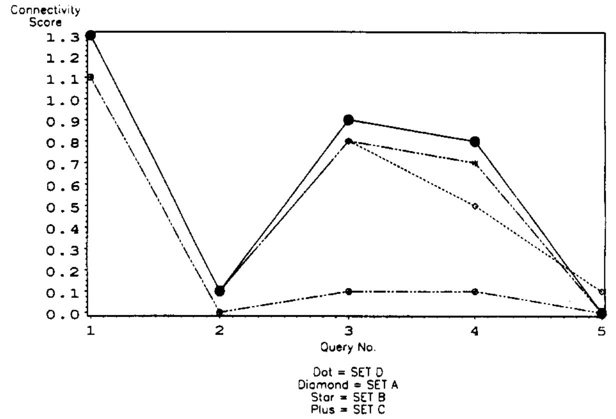


Figure 12. Connectivity score for key features in top 20 retrievals for the pilot-5 query set using SET-A, SET-B, SET-C, and SET-D feature sets.

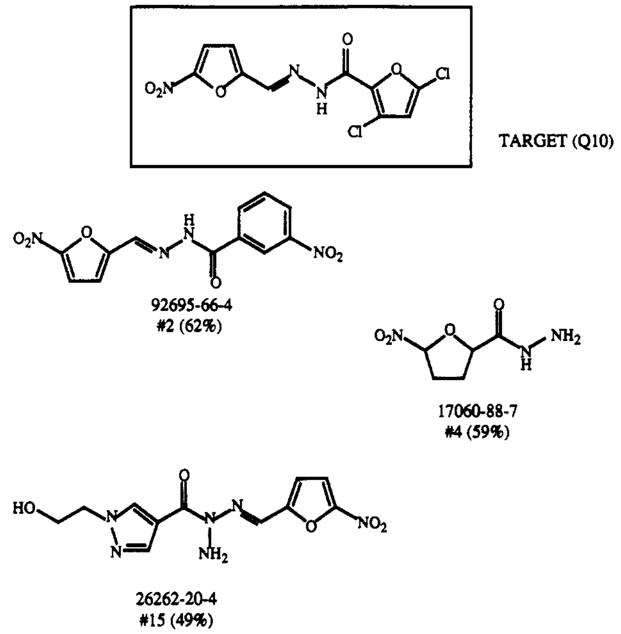


Figure 13. Some sample fragment-based retrievals for query no. 10 using the SET-A feature set.

Although SET-A with over 1300 AA and AS screens is a significant portion of the total feature set, SET-D, the additional screens in SET-D can play an important part in the retrieval ranking. Figure 13 illustrates some of the top ranked retrievals for query no. 10 using the SET-A feature set. The second, fourth, and fifteenth ranked retrievals are shown. Intuitively, it would seem that the ordering of the fifteenth and the fourth retrievals should be reversed. The fifteenth retrieval is somewhat larger than the query structure, but it has two ring systems with the pyrazole system being a reasonable surrogate for a furan ring. The fourth retrieval is approximately half the size of the target structure. The problem with respect to the SET-A feature set seems to be the symmetry of similar groups in the target, and even though there are a significant number of AA and AS screens generated (107), few of them reflect two or more occurrence counts and, thus, the screens are found in approximately half of the structure.

The top 20 SET-A retrievals for query no. 10 were reranked using the SET-D feature set. The fourth SET-A retrieval was ranked last; the fifteenth was ranked twelveth. This SET-D ranking seems to be more appropriate and is probably due to the presence of classes such as atom counts and element

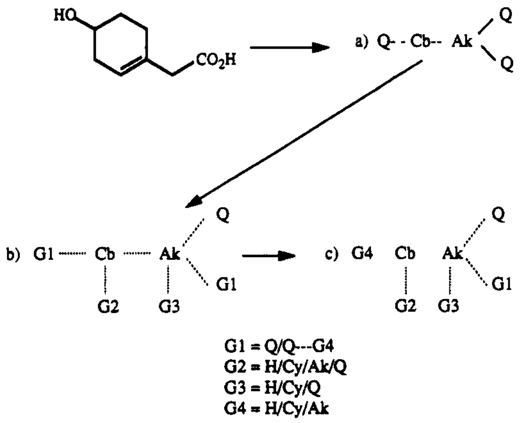


Figure 14. Basic generic structures used in searching the STN 30K answer set.

composition which help specify the overall size and composition.

Fragment-Based Recall Assessment. A recall assessment of fragment-based similarity searching was performed using the SET-A feature set and the full set of 20 queries. First, the fragment-based searches were executed against the 30K file on the modeler and the top 350 retrievals from each search was transferred to STN as a saved answer set. Generic group structures corresponding to the queries were then searched for in a subset search of the full 30K STN answer file. The results of these searches were NOTed with the corresponding 350 answer set from modeler, and the structures in the resultant sets were examined manually.

The 350 top ranked answers from the fragment-based searches against the 30K file roughly extrapolates to 100K answers from a full CAS Registry File search, assuming, of course, a linear relationship. Thus, all relevant answers should be in these first 350 fragment-based answers. Searching the full 30K answer set on STN with generic structure queries should capture most of the relevant answers, at least from our perception of relevancy. The resultant set after the Boolean NOT operation should contain any recall features; i.e., they would not be in the first "100K" answers relative to a full Registry File search.

Three basic generic structures of different degrees of specificity were used in searching the 30K file on STN; these are illustrated in Figure 14 for the hydroxycyclohexeneacetic acid query. The most specific one is the connected hierarchical generic group structure, "a", executed as a "closed substructure search" (CSS), i.e., no further substitution allowed. In the "b" structure, also executed as a CSS, one substituent is permitted on each of the framework generic groups. For example, the G2 on the Cb (carbocycle) is H (i.e., no substituent), Cy, Ak, or Q (i.e., a substituent). The "c" structure is the broadest one and has disconnected fragments, i.e., in the retrievals there may be intervening groups between those shown in the query. The type of generic group structure to be used for a query was determined by trial and error. The goal was to obtain at least a hundred answers for analysis *after* the Boolean NOT operation between the answer files. In several cases, even structure "c" was too specific and only some of the disconnected fragments were used in searching to obtain a final answer set of the appropriate size.

For the 2D query set, approximately 2K answers were examined to determine if any recall failures were present. Even with a liberal view of relevancy, no definitive recall failures were found. Thus, recall is not likely to be a problem for fragment-based searching with a general STN screen feature set, at least with respect to the first 100K ranked answers.

Secondary Fragment-Based Searching and Reranking. A general feature set such as SET-D should provide for "complete" recall and a good overall ranking for a 100K answer set from a FULL CAS Registry File search on STN. However, the perception of a good ranking may vary considerably from user to user. If, for example, a user is interested in the 26262-20-4 substance relative to query no. 10 (see Figure 13), this substance could be a few thousand "deep" in a 100K answer set, assuming a linear extrapolation of the results of the 30K file SET-D search (where the 26262-20-4 substance was ranked no. 12).

One way to accomplish a reranking of an answer set is via secondary fragment-based searching on the results of a general similarity search using specific screen class feature sets. For example, in a query 10 search using the AC feature set against the top 20 retrievals for the same query using the SET-A feature set (see Figure 13), the 26262-20-4 substance was ranked second; using the TR feature set it was ranked first; and using the EC feature set it was ranked fourth. Such secondary searching gives a "size", a "ring" and an "atom flavor" similarity view, respectively.

Feature sets which do not have features with certain characteristics may be useful in collecting together substances with useful variations on the missing characteristic. For example, in the top 20 answers using SET-D for the hydroxycyclohexeneacetic acid query (query no. 1; see Figure 6), there were only two derivatives of the carboxylic acid functional group: two closely related carboxy esters. However, when one executes a CS feature set search against the top 350 SET-D answers, a more diverse set of acid derivatives is obtained, including a $-C(=O)NH_2$, $-C(=NH)NH_2$, and $-C(=NH)S-$. Presumably, this is due to the lack of atom flavoring in the CS screens.

Individual feature weighing is a useful way to accomplish a reranking on a local basis by "requiring" via the weighting that retrievals have certain query features, i.e., that substances with those features be ranked near the top. For example, in the top ranked 20 retrievals using the SET-D feature set for query no. 1, there are nine substances with the CO_2H group. To test the impact of individual feature weighting, four out of the 75 SET-D screen features for the query were assigned a weight of 5 with the other 71 screen features having a default weight of 1. The weighted features were augmented atoms associated with the carboxylic acid group: (C-C-O-O), C-O-O), (C-4O-4O), and (C O O or C N S). The first 350 SET-D answers were then searched with the weighted query. In the top 20 results from weighted search, all the substances, except one, had a $-CO_2H$ group; the exception had a carboxylate anion. Thus, the individual feature weighting can be used to rerank an answer set on a local basis according to user specified important and necessary query features.

Another important use envisaged for fragment-based similarity searching is the ranking of an answer set from a substructure (inclusion-match) search. A complete structure close to the substructure query or a prototypical answer from the search would be used as the target or query structure for the similarity subset search on the substructure answer set.

Connectivity-Based Searching. The current STN connectivity-based search capabilities can be used to provide more precise global or local similarity definitions on answer sets derived from one or more fragment-based similarity searches. Large answer sets can be significantly reduced via such

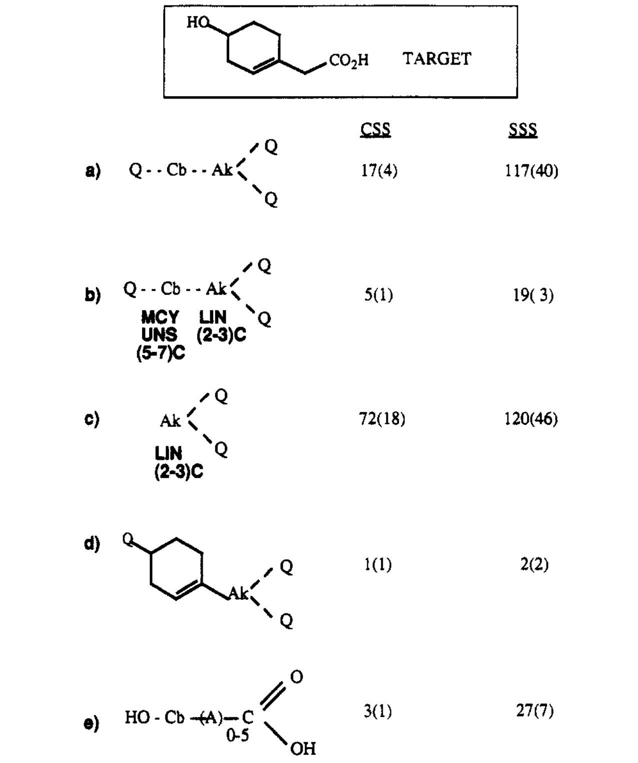
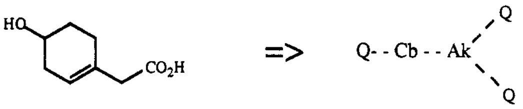


Figure 15. Sample queries for connectivity-based similarity searching of SET-A fragment-based search results for query no. 1.

searching. Figure 15 gives a set of connectivity-based queries used to refine the first 350 ranked SET-A results for the hydroxycyclohexeneacetic acid query (query no. 1). The corresponding number of retrievals from searches using these queries against the 350 subset is given in the columns on the right of Figure 15. The CSS column gives the number of retrievals obtained when the query is executed as a closed substructure search, i.e., no further substitution. The SSS column gives the number of retrievals obtained when substitution is permitted; i.e., the real-atoms corresponding to the generic group may be attached to other real atoms. The numbers of parentheses are the results of a subsequent subset search with a substructure query consisting of a carbon–carbon ring double bond fragment. The purpose of this additional search is to ensure the presence of a non-aromatic ring, i.e., to eliminate substances where the carbocycle is a benzene ring.

The "a" query is the corresponding hierarchical generic group structure. It reduced the answer set from 350 to 17 for the CSS search which is further reduced to 4 with a subsequent search for a ring double bond, i.e., an 87-fold reduction. The "b" query is the "a" query qualified by generic group categories and attributes. The carbocycle must be a monocycle, be unsaturated, and have from five to seven carbon atoms. The carbon chain group, Ak, must be linear and must have from two to three carbons. The "b" query is very precise since after the ring double bond restriction only one answer is obtained, i.e., the target or query structure.

The "c" query illustrates the use of connectivity-based searching for a portion of the target, i.e., a local similarity. The query specifies a functionality that includes a carboxylic acid or a carboxylic acid derivative. If the results of a CSS search using this query are reranked by a subsequent fragment-based search, the final results should be similar to those of the individual feature weighting results discussed above. However, the generic group structural fragment results will also include acid derivatives such as –C(=O)NH2, –C(=O)Cl, etc., which should be relevant to most users. For a CSS search with this generic functional group, the answer set was reduced from 350 to 72. A SET-D fragment-based search could be used to create a ranking for this reduced set containing substances with the carboxylic acid or acid derivative.



3 Q + Cb + Ak	1 each (3)
Q -- Cb + Cb -- Ak + Ak -- Q	2 each (6)
Q - - Cb - - Ak + Cb - - Ak - - Q + Q - - Ak - - Q	3 each (9)
Q - - Cb - - Ak - - Q	10 each (10)
Q - - Cb - - Ak (- - Q)2	15 each (15)
	Maximum Score (43)

Figure 16. Calculation of a possible connectivity-based similarity metric.

Occasionally it is desirable to "freeze" a portion of the target structure, i.e., require that it must be present is all answers. Query "d" species the requirement of a cyclohexene fragment with a positional relationship for the two generic group substituents. In STN structure handling, any valid combination of generic group and real-atom structural fragments is permitted with respect to subset searching. Query "d" is a very precise search yielding only one CSS answer, the target, in a subset search on the initial 350 answer set.

Another type of specification that is useful in similarity searching is the location of two groups of atoms (or surrogates for such groups) within a certain number of atoms or bonds from each other. Query "e" is a mixture of real-atom and generic group fragments in which the carbocycle group must be within zero to five real-atoms (A for "any" non-hydrogen real-atom) from a carboxy group. This query is also very specific with only three CSS answers out of the 350 subset.

In addition to generic group structures, other STN features are also available to help formulate precise similarity definitions. A subset search using a substructure query was already mentioned. Subset searching using a query that is an expression of screen numbers can also be useful. For example, screen expressions can be developed to specify the range of non-hydrogen atoms in the answers using the AC screens. Similarly, the ring analysis data can be useful for additional ring information. For example, a subset search using the "6,6" ring size could further qualify the answers derived from a "Cb, polycycle" specification.

The generic group and related structures illustrated in Figure 15 could be built for subset connectivity-based similarity searching using existing STN capabilities. Fortunately, such queries are fairly small and should be easy to construct. However, eventual automatic construction of several generic group structure queries of varying levels of specificity would seem to be appropriate, especially for novice users. Such structures might include the basic structures used for the recall assessment experiment (see Figure 14).

After execution of a connectivity-based subset search a reranking of the resulting answer set may be needed if it is still of substantial size. As mentioned above, the reranking could be accomplished via subsequent fragment-based similarity search. Another possibility would be to use a complementary connectivity-based measure for the ranking. A possible scheme for a connectivity-based measure is illustrated in Figure 16. This would involve the decomposition of the hierarchical generic group structures for the query and the file substances in the fragment-based search answer set into all one-group, two-group, three-group, etc., fragments. The fragment lists for the query and each file substance would be compared. For every one-group query fragment found in the one-group list of the file substance, the overall score would be incremented by the value assigned to a one-group match (one point is illustrated); for every two-group fragment found, the score would be incremented by the two-group value, etc. The net result of this process is that the greater the connectivity of a file substance relative to the query, the greater the overall score. This scoring could be incorporated into a Tanimoto framework, but this might not be needed since the size differential between the query and file substance is taken into account in the fragment-based searching, to a large extent. This connectivity-based measure seemed to be effective in a manual calculation and comparison of a query no. 10 (see Figure 13) and several top ranked SET-A answers. However, automation of procedure and more extensive testing are needed to obtain definitive results.

SUMMARY AND CONCLUSIONS

This study has assessed the potential for global and local 2D structural similarity searching on STN structure files using the existing substructure search screens and structure handling capabilities.

The STN substructure screen set contains a rich set of structural fragments in the individual screen classes and in class combinations which provide effective, multiple views of 2D structural similarity. These similarity views would be obtained by fragment-based searching using the Tanimoto similarity coefficient. We have found that the full set of screens is an effective "general" set of screen fragments for the determination of 2D structural similarity. This set is a prime candidate for the default set for similarity searching. The full set had the best overall relevancy among several class combinations we have examined. Also, assuming ranked answer sets of up to 100K substances from a similarity search, the recall for a target substance should be complete, using any one of the various general set candidates. We are continuing to explore manually selected subsets of the STN substructure screen set for use as a general set to see if even greater performance can be obtained.

The individual STN screen classes provide different similarity views. For example, the bond sequence class gives a "bonding" view and the hydrogen augmented atoms and the twin augmented atoms a "substitution" view. An anticipated use for the individual classes is in a subset search of an answer set so as to obtain a reranking of the answers according to the similarity view of the individual class. Also, an answer set could be reranked to provide a local similarity view by individual screen fragment weighting. This type of weighting can be used to require retrievals to have a structural portion that is very similar or identical to one in the target structure. The general and/or individual class sets should also be useful in ranking a substructure search answer set.

Existing STN structure handling capabilities provide a precise, connectivity-based similarity view. There is a considerable amount of flexibility in obtaining a precise global or local similarity view due to the specific–generic continuum in STN structure representation. Query structures can be constructed consisting solely of connected generic groups (e.g., carbocycle) which would require a connected framework in the retrieved similar substances. The generic group structure nodes can be qualified with categories and attributes for additional precision. If desirable, real-atom fragments can be used in the query structures along with generic groups to require that such real-atom fragments be present in all retrievals. Also, a local similarity view can be accomplished by using a generic structure corresponding to a portion of the target structure such as a functional group. Such a query structure would require that retrieved structures have a real-atom moiety corresponding to the generic fragment, for example, a carboxylic acid along with surrogates such as an amide, acid chloride, etc.

A typical interaction scenario anticipated for 2D structural similarity searching on STN structure files might involve (1) recall or creation of the target structure, (2) execution of a sample followed by a full file fragment-based similarity search using the general (default) screen set to create a ranked set of up to 100K answers, (3) browsing the ranked answer set and, if appropriate, reranking the answers via a subset similarity search using individual screen classes or via individual feature weighting on a selected portion of the target structure, (4) establishing a connectivity framework among the various atom groups or their surrogates in the target structure on a global or local basis via subset searching of the fragment-based answers with a generic group or a mixed generic group/real-atom structure, and (5) reranking the resulting, reduced answer set with a subsequent fragment-based similarity subset search or, perhaps, with a new connectivity-based similarity measure.

The first three steps in the scenario should be fairly easy to execute even by a novice user, especially if the target structure can be obtained via a Registry Number recall, i.e., no structure building. The secondary searching specified in steps 3–5 would be optional. Step 4 would require that a user be familiar with existing STN query structure building. Several types of generic group structures of various degrees of specificity could be automatically generated from the target structure to support novice users in this task.

We conclude that coupling of fragment-based similarity searching using existing STN substructure search screens with connectivity-based searching using query structures containing generic groups is a viable concept which allows for a significant amount of flexibility in the user's definition of 2D structural similarity, including multiple similarity views on either a global or local basis.

We plan to continue to explore aspects of 2D structural similarity as well as size and shape similarity, chemical similarity, and the clustering of substances in order to meet a variety of similarity searching and answer file manipulation needs for users.

ACKNOWLEDGMENT

We would like to thank the following members of the CAS Research staff for their assistance in this project: T. E. Bangert, K. P. Cross, and G. G. Vander Stouw.

REFERENCES AND NOTES

(1) Lawson, A. J. Organic Reaction Similarity in Information Processing. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 675–9.

(2) Gasteiger, J.; Ihlenfeldt, W.-D.; Fick, R.; Rose, J. R. Similarity Concepts for the Planning of Organic Reactions and Syntheses. *J. Chem. Inf. Comput. Sci.*, **1992**, *32*, 700–12.

(3) Sheridan, R. P.; Vekataraghavan, R. New Methods in Computer-Aided Drug Design *Acc. Chem. Res.* **1987**, *20*, 322–9.

(4) Fisanick, W.; Cross, K. P.; Rusinko, A., III. Characteristics of Computer-Generated 3D and Related Molecular Property Data for CAS Registry Substances. *Tetrahedron* **1990**, *3* (6C), 635.

(5) Gieschen, D.; Rusinko, A., III; Vander Stouw, G. G.; Fisanick, W.; Lillie, D. H.; Reams, N. Automatic Screening of Large Chemical

Structure Databases for Potential Biological Activity; *Abstracts of the QSAR 92 Conference*, Duluth, MN; 1992.

(6) Fisanick, W.; Cross, K. P.; Forman, J. C.; Rusinko, A. III. Experimental System for Similarity and 3D Searching of CAS Registry Substances. 1. 3D Substructure Searching. *J. Chem. Inf. Comput. Sci.*, **1993**, *33*, 548–9.

(7) Fisanick, W.; Cross, K. P.; Rusinko, A., III. Similarity Searching on CAS Registry Substances. 1. Global Molecular Property and Generic Atom Triangle Geometric Searching. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 664–74.

(8) Johnson, M.; Naim, M.; Nicholson, V.; Tsai, C. C. Comparing the Substructure Metric to Some Fragment-Based Measures of Intermolecular Structural Similarity. *Pharmacochem. Libr.* (*QSAR Drug Des. Toxicol.*) **1987**, *10*, 67–9.

(9) Willett, P.; Winterman, V. A. Comparison of Some Measures for the Determination of Intermolecular Structural Similarity: Measures of Intermolecular Structural Similarity. *Quantum Struct.-Act. Relat.* **1986**, *5* (1), 18–25.

(10) Willett, P. *Similarity and Clustering in Chemical Information Systems*; Research Studies: Letchworth, U.K., 1987.

(11) Johnson, M. A., Maggiora, G. M., Eds. *Concepts of Molecular Similarity Analysis*; Wiley-Interscience: New York, 1990.

(12) Clements, J.; Hicks, M. G.; Sunkel, J. The 1992 Beilstein Workshop on Similarity in Organic Chemistry. *J. Chem. Inf. Comput. Sci.*, **1992**, *32*, 577.

(13) Takahashi, Y.; Sukekawa, M.; Sasaki, S. Automatic Identification of Molecular Similarity Using Reduced-Graph Representation of Chemical Structure. *J. Chem. Inf. Comput. Sci.*, **1992**, *32*, 639–44.

(14) Hagadone, T. R. Molecular Substructure Similarity Searching: Efficient Retrieval in Two-Dimensional Structure Databases. *J. Chem. Inf. Comput. Sci.*, **1992**, *32*, 515–21.

(15) 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.

(16) 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 106–29.

(17) 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.

(18) Fisanick, W. Storage and Retrieval of Generic Chemical Structure Representations. U. S. Patent 4,642,762, Feb 10, 1987.

(19) Ebe, T.; Sanderson, K. A.; Wilson, P. S. The Chemical Abstracts Service Genetic Chemical (Markush) Structure Storage and Retrieval Capability. 2. The MARPAT File. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 31–6.

(20) The SAS System is available from the SAS Institute, Inc., SAS Circle, P. O. Box 8000, Cary, NC 27512–8000.