

Structural Similarity Searching Using Descriptors Developed for Structure-Activity Relationship Studies

Philip N. Judson

Heather Lea, Bland Hill, Norwood, Harrogate HG3 1TE, U.K.

Received June 15, 1992

A similarity searching system is described, based on the use of descriptors developed for the REX pharmacophore recognition system.⁶ The use of these descriptors, which included hydrogen, lone electron pairs, multiple bonds, and the concept of generic atom types, together with hierarchical application of search criteria gave answer sets and rankings that were reasonable from the point of view of a chemist interested in the biological activities of compounds.

INTRODUCTION

A point to be taken into account in the design of similarity searching systems is that similarity is context-dependent; the answer expected to a query depends upon the criteria for similarity that the user applies, consciously or subconsciously. This contrasts with a substructure search system, for example, which a user expects to return all compounds containing specifically the substructure entered as a query, and only those—an expectation that it is possible to meet with few exceptions. Many types of descriptors have been used to provide similarity searching systems suited to different needs.¹


Chemical structure database management systems such as MACCS² use descriptors derived from substructure-search screens for similarity searching. Although these descriptors are remote from the ways in which a chemist probably thinks about similarity, their use can give acceptable results, and correlations have been reported between the similarity rankings obtained and biological activity.³ However, results are often puzzling to users, and there are many situations in which the method fails.¹

A process dependent on the recognition of similarity is the analysis of structures of compounds with a common biological activity, to discover what features are responsible for the activity. Several systems have been described which aim to do this and which use the knowledge so gained to make predictions about novel compounds. For example, TOPKAT⁴ uses a large set of predefined descriptors in an approach somewhat analogous to the use of substructure-search screens for similarity searching. CASETOX⁵ uses complete sets of linear fragments generated from all the structures included in each analysis. REX⁶ uses more generalized linear descriptors and goes on to build the more complex fragments known as pharmacophores.

Carhardt, Smith, and Venkataraghavan⁷⁻⁹ have described successful similarity searching based on linear descriptors, "atom pairs", like those generated by REX. However, these atom pairs did not include some of the features included in REX descriptors.

This paper describes the application of descriptors developed for REX to similarity searching, to provide a system for a user who thinks primarily in terms of the features of a molecule responsible for its biological activity. In practice, the use of these descriptors has been found to give results that are consistent with the expectations of chemists in general.

Table I. Set of Links for 2-Bromocyclopropylamine



terminator 1	terminator 2	link length	no. of occurrences
Br	C	2	2
Br	C	3	2
Br	N	3	1
Br	N	4	1
Br	lone pair	4	1
Br	lone pair	5	1
Br	H	4	2
Br	H	5	2
N	C	2	2
N	C	3	2
H	H	2	1
H	C	2	2
H	C	3	4
H	C	4	4
H	lone pair	2	2
lone pair	C	2	1
lone pair	C	3	2
lone pair	C	4	2
halogen	C	2	2
halogen	C	3	2
halogen	N	3	1
halogen	N	4	1
halogen	lone pair	4	1
halogen	lone pair	5	1
halogen	H	4	2
halogen	H	5	2

CHOICE OF DESCRIPTORS

A REX descriptor, which will be called a "link", is a pair of "terminators" and the distance between them, currently expressed as the number of bonds. A terminator may be an atom, a lone pair, or a bond. All such links are determined for every structure included in a set under study, subject to certain restrictions described below. The set is not restricted to links for the shortest paths between pairs of terminators—all paths are included. If a structure contains more than one fragment (for example, if it is a salt), all the links for all the fragments are computed.

Table I shows the set of terminators for 2-bromocyclopropylamine, as an example. For the work described in this paper, 10 elements were used individually as terminators—H, C, N, P, O, S, F, Br, Cl, and I. These elements were chosen because of their biological importance. Other elements were classed together automatically by the system as a single terminator, "M".

Links having two carbon terminators are normally excluded from REX analyses. This reduces considerably the total number of links associated with each structure, with minimal loss of information. Most structures of biological interest contain at least one heteroatom or center of unsaturation, which contributes to binding at the site of action. All carbon atoms in these structures are thus taken into account by being in links to non-carbon terminators. For an analysis of saturated hydrocarbons it would be necessary to include the links between pairs of carbon terminators.

Individual hydrogen atoms are included as terminators only if they are attached to specified atom types (which were nitrogen, oxygen, and sulfur for the work described in this paper) at specified valency levels. Lone pairs can be defined as terminators, and for the work described in this paper those on trivalent nitrogen, divalent oxygen, and di- and tetravalent sulfur atoms were included. For the work described in this paper, all types of multiple bond were classed together as a single terminator type.

The types of atoms and bonds connecting two terminators are not taken into account—only the types of the terminators and the length of the link. This description recognizes that when parts of a molecule bind to specific points in a site of biological action, it is the distance between the binding points that is important, and not the nature of the connections between them. Although the distances that are important biologically are 3D distances, the use of topological distance can place satisfactory constraints on 3D distances for this type of analysis. In practice, 3D distance ranges are of interest, rather than exact values, to allow for flexibility. The 3D distance between a pair of terminators separated by a certain number of bonds can only fall within a certain range. Thus, the use of topological distance is one way of recognizing links having 3D lengths within the same range. Although the 3D range for a long link is large, there are two reasons why this may not be a problem in practice. The first is that if two structures have the same, high degree of flexibility, that is a similarity between them in itself. The second is that any structure containing a long link must also contain a large number of shorter links, which impose tighter constraints and influence the overall measure of similarity between the structures. For the work described in this paper, link lengths were restricted to the range 2–20.

The links associated with each molecule in a reference set are compared with those associated with the given query structure to rank the members of the reference set in order of similarity to the query. Various ways of computing measures of similarity for ranking have been discussed.^{3,10} The results reported in this paper were based very simply on the difference between the number of features in common between two structures and the number of features found only in one structure or the other, i.e.

$$S = c - u$$

where S is the similarity score, c is the number of descriptors in common, and u is the number of descriptors unique to one or the other structure. This was found to be an effective measure for similarity when it was used in conjunction with the hierarchy described below, although there would be limitations to its use for other purposes, such as clustering.

GENERIC TERMINATOR TYPES

When scientists compare structures, they frequently take account of similarity at an atomic level and may allow this to have an overriding importance. Figure 1 shows a set of compounds recognized as similar to the target, 1. It is likely

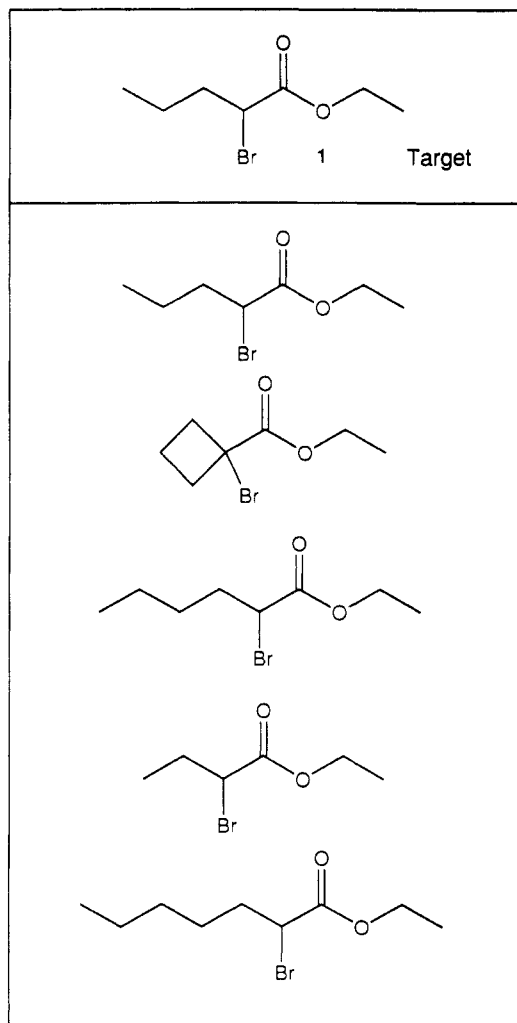


Figure 1. Results of a search for compounds similar to ethyl 2-bromopentanoate.

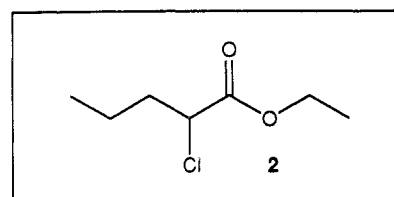


Figure 2. Ethyl 2-chloropentanoate.

that a chemist would class the chloro analogue 2 in Figure 2 as more similar to the target than any of the other compounds in Figure 1. In order to mimic this kind of decision, the similarity system has been designed to recognize generic classes. For the work described in this paper, the halogens were defined as a generic class—in this case causing the system to rank structure 2 as second in the list.

MULTIPLE OCCURRENCES OF LINKS

Figure 3 again illustrates rankings that take into account the generic halogen class and illustrates another potential problem. Which of the two rankings of the dibromo compounds 3 and 4 is more appropriate? This is a case which might be described as a matter of opinion. The author prefers the ranking in the right-hand column in Figure 3, in which the geminal dibromo compound 4 takes precedence.

Referring to Figure 3, consider the specific link of a bromine atom and the double bond of the carbonyl group through two intervening bonds. The link occurs once in the target and

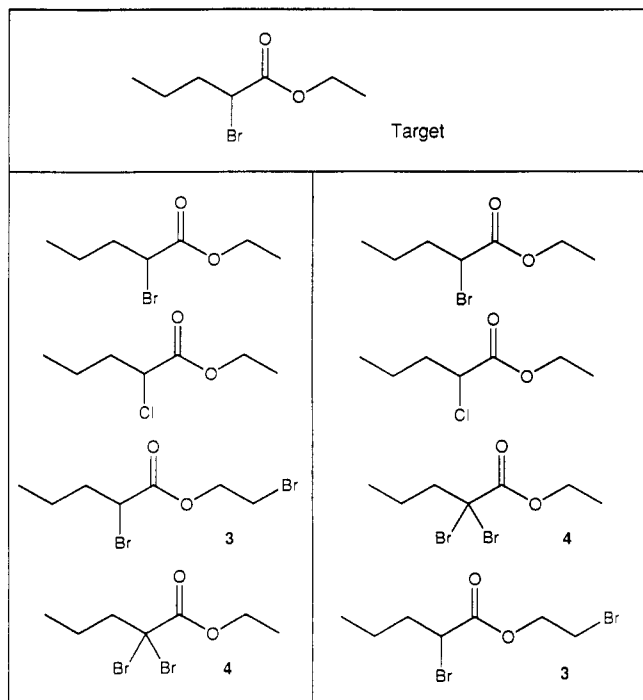


Figure 3. Alternatives for the ordering of structures containing different numbers of occurrences of the same links.

once in 3. It occurs twice in 4. Consider the link of a bromine atom to the double bond through four intervening bonds. This link occurs once in 3, and it is absent from the target and 4. Looking at all the possible links in the three compounds reveals that while 3 contains many new bromine-terminated links (as well as one co-incidental duplication—a bromine atom separated from an oxygen atom by three bonds), 4 contains only one new link (two bromine atoms separated by two bonds) but contains duplicates of all of the bromine-terminated links in the target. In general, taking into account the relative number of occurrences in each structure of each link common to a pair of structures usually produces rankings like the one in the left-hand column of Figure 3; taking account only of the number of common links, with no penalty for differences in the number of occurrences in each structure produces rankings like the one the right-hand column of Figure 3.

SORTING HIERARCHY

To take full advantage of the observations made above about generic atom classes and multiple occurrences of links, the similarity system ranks structures according to a hierarchy. The question of whether or not to take account of the number of occurrences of a descriptor in each structure has been discussed before.³ In the system described in this paper, both options are used. Primary sorting is based only on the number of link-types in common between structures. In addition, at this stage, the members of each generic atom class are treated as equivalent. If structures have equal ranking on this basis, a secondary sorting takes account of the number of occurrences of each link-type in each structure. Finally, a third level of sorting takes account of the specific element types belonging to generic classes that are found in the structures.

IMPLEMENTATION OF THE SIMILARITY SEARCH SYSTEM

The current version of the similarity search system runs on a VAX computer under VMS. Connection tables (derived, for example, from MACCS MOLfiles) are analyzed, and

information about the links associated with each structure is stored in an intermediate, archive file. This process is potentially slow, since it involves a back-tracking algorithm, but needs only to be carried out once for a given set of compounds (information about additional compounds can be added later, without the need to reprocess compounds already included in the archive).

Subsequently, the user can select as the target a compound from the main archive file, or use a different one, since, even with backtracking, the initial analysis is sufficiently quick to deal with a single compound virtually instantaneously. Currently, the output after analysis and ranking is a list of compound identification codes. The list can be used as a MACCS listfile, for example, so that the structures can be viewed in order of ranking.

PLANNED ENHANCEMENTS

For the work described in this paper, only one generic class of atoms (the halogens) was used. The system has been designed to support multiple generic classes and additional classes will be added. The use of generic classes is not restricted to groupings of related elements. Hydrogen atoms in different environments can be treated as discrete members of the hydrogen class, for example, and lone pairs on atoms in differing environments are distinguished in REX, taking account of both the valency state and the coordination level of an atom. Multiple bonds were represented by a single terminator for the work described in this paper. This terminator may also be declared to be a generic class, containing terminators for each of the main multiple bond types—double, triple, and aromatic.

CONCLUSION

This paper describes a further step in the development of similarity search methods based on fragment descriptors. Results consistent with the expectations of a chemist interested in biological activity have been obtained for a variety of structure types, and the examples which follow show that the system can recognize similarity between superficially different structures such as different heterocycles. At the same time, by creating a full set of descriptors dynamically for every structure, the system avoids the tendency of systems based on predefined fragments to rank markedly-differing structures together if they contain the same fragments in substantially different spatial relationships. Further experiments will be necessary to determine whether there is good correlation in practice between the similarity rankings given by the system, and biological activity.

EXAMPLES

Two test sets of structures were created from the MACCS version of the *Fine Chemicals Directory*.¹¹ The first contained a wide range of 505 structure types chosen arbitrarily. The second contained 182 structures, made up of a variety of aromatic heterocyclic compounds, selected at random, together with a few benzene derivatives.

In each example described in this paper, the first few structures from the ranked list are shown. It is not practical to list here all 505 members of each set, and it would not be useful so to do since below some cutoff point the structures in a random set cease to be similar by any normal criteria. Of the searches tried, there were several not reported here in which the target structure chosen happened to be unique in the random set, making the search pointless. However, there

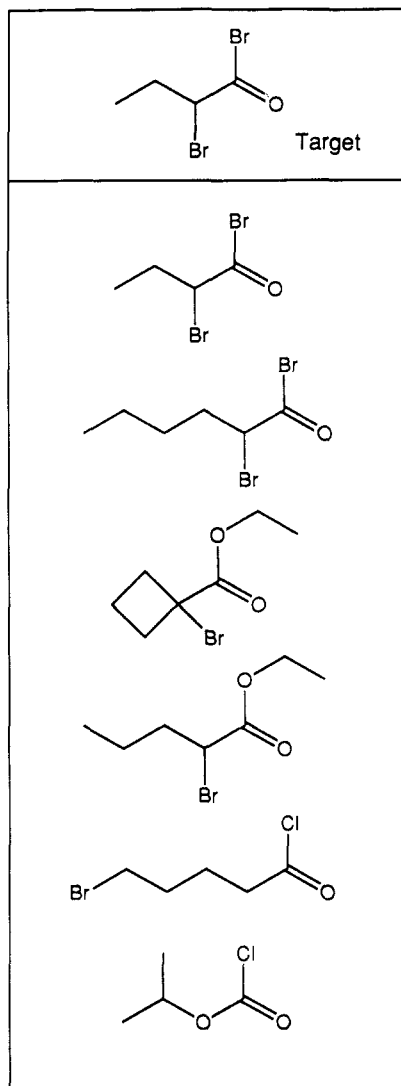


Figure 4. Results of a search for compounds similar to 2-bromobutanoyl bromide.

were no cases of search results that were inappropriate (for example, there were no cases seen in which a structure that a chemist would think was similar failed to rise to the top of the ranked list or in which a noticeably dissimilar structure ranked above similar ones). I would be pleased to supply full lists of structures to the reader on request.

Results obtained by searching in the first set were as follows. Ethyl 2-bromopentanoate as target gave the list shown in Figure 1. Note that, as in most cases, the exact match was ranked first. Since all the links present in every structure are determined dynamically, the system is normally able to assign highest ranking to the exact match. A specific, predictable exception is described below with regard to the structures in Figure 8. The ranking of the second compound in the list in Figure 1 would be unlikely to be found with other systems, since most attach weight to the presence of a ring. Seen from a biological point of view, the ranking is not unreasonable—the two structures differ only by the creation of a single bond with the loss of two hydrogen atoms, to make a slight modification to a small, lipophilic fragment. The third and fourth structures contain one more, and one less, CH_2 fragment than the target, and the fifth contains two more. Rankings cannot be exact within one or two places in any system dealing with an imprecise notion like similarity, and no particular significance is attached to the ordering of these structures. It is more important that the system is able to select the five, closely-

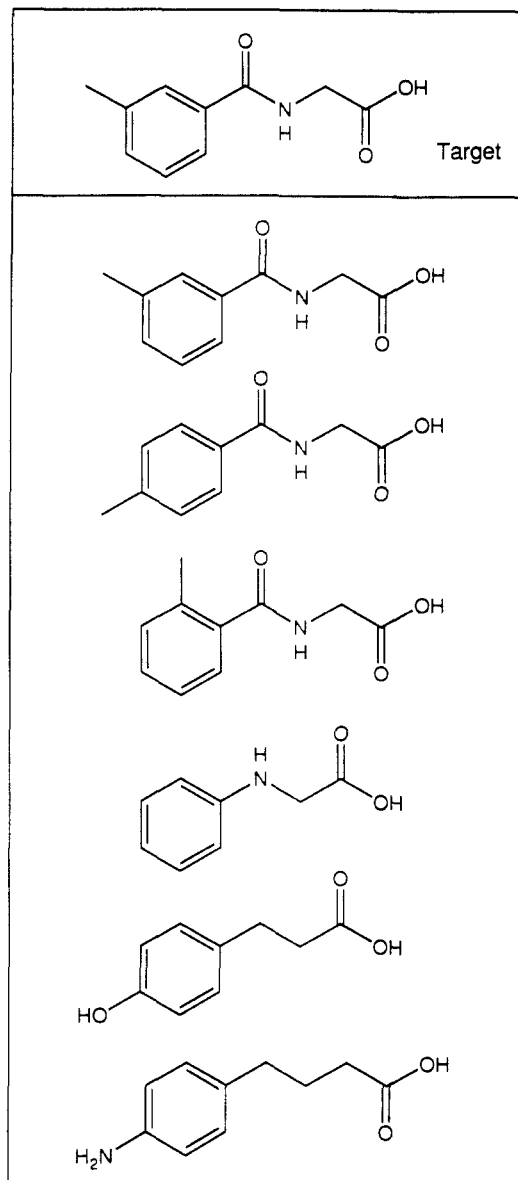


Figure 5. Results of a search for compounds similar to 3-methylbenzoylglycine.

related structures from a widely-diverse set. Nevertheless, it is encouraging that the ordering of the list shows a generally logical trend.

The listings shown in Figure 3 were created from a small subset of structures for the purpose of illustration in the discussion above, and they will not be discussed in any more detail here.

2-Bromobutanoyl bromide as target gave the list shown in Figure 4. The exact match was ranked first, followed by the only other 2-bromocarboxylic acid bromide in the set. It is debatable whether the third and fourth structures (the cyclobutanecarboxylate and the pentanoate) might more appropriately be listed in reverse order. The next structure in the list was an acid halide, like the target compound, but it lacked substitution at the 2-position. The last structure in the list, a chloroformate, was still related to the target structure, but was more remotely similar.

The next structure chosen as target was 3-methylbenzoylglycine, which gave the list shown in Figure 5. Following the exact match, the 4- and 2-substituted analogues were found (the ranking of these two was actually equal, and the order of their listing is arbitrary). The next structure, *N*-phenyl-

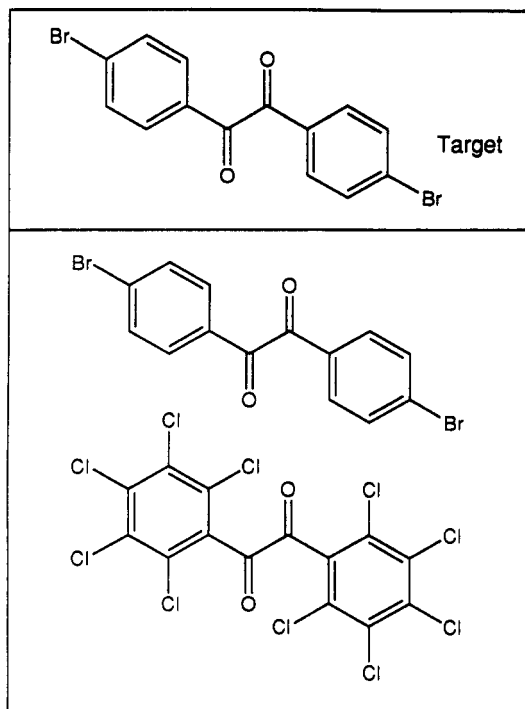


Figure 6. Results of a search for compounds similar to 4,4'-dibromobenzil.

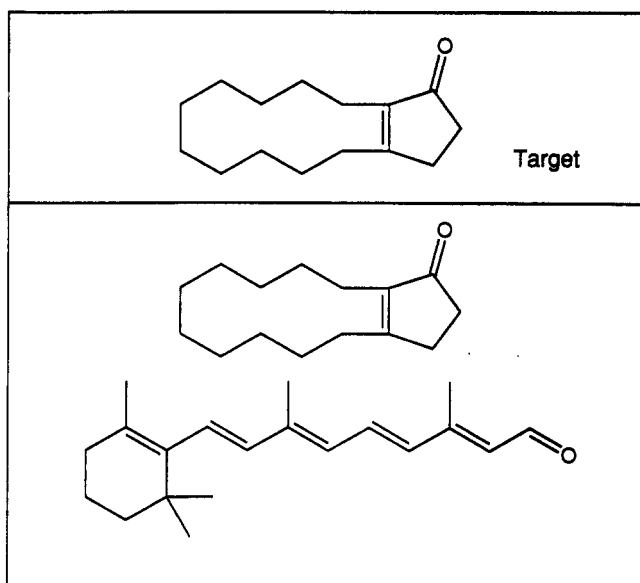


Figure 7. Results of a search finding a compound not closely related to the target but having significant features in common with it.

glycine, marks a significant step in similarity rating, but it is a valid selection from the wide set of alternatives. The last two compounds in the list are more remote, but both of them contain a benzene ring at approximately the same distance from a carboxylic acid group as in the target, and their selection is therefore not unreasonable.

The main set contained only one compound that could be said to be similar to 4,4'-dibromobenzil, the target in Figure 6. The system found the decachloro analogue successfully.

There were really no compounds (apart from the exact match) that would normally be regarded as similar to the target shown in Figure 7, but the next most similar structure has been included in these examples to show that the system can find remote similarity. The compounds may be considered to have some similarity in a biological context, since they both contain an enone attached to a large, lipophilic fragment.

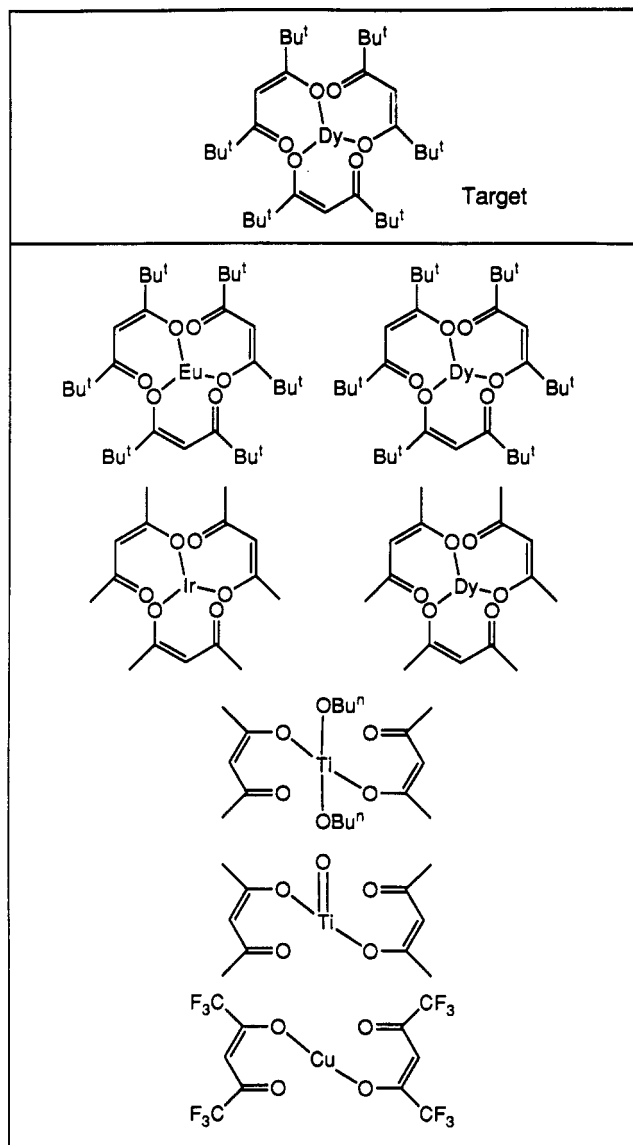


Figure 8. Results of a search for compounds similar to an organometallic complex.

A very different target was the shift reagent in Figure 8. In the experiments described in this paper, no distinction was made between different metals; so the europium analogue of the target was ranked equally with the exact match; and two structures were also ranked together in second place (as mentioned in the discussion of choice of descriptors above, all elements not declared individually for use as terminators are classed as equivalent by the system).

The remaining examples were obtained from searches of the set of heterocyclic structures. Using an acridine as target gave the results shown in Figure 9. The second structure in the answer set is an example of a compound containing more than one fragment in the MACCS database, which has been successfully recognized as a close relative of the target structure. As stated above, too much significance should not be attached to exact rankings. However, it is worth noting that the ranking of the third and fourth compounds in the list would be reversed if oxygen and nitrogen were declared as a generic pair. The compound in position 3 might also be considered to rank below the compound in position 5 from a biological point of view. 3 is superficially more similar to the target than 5, because it has a more similar ring system, but 5 contains an amino group and a lone pair associated with a pyridine nitrogen, separated by the same distance as the same

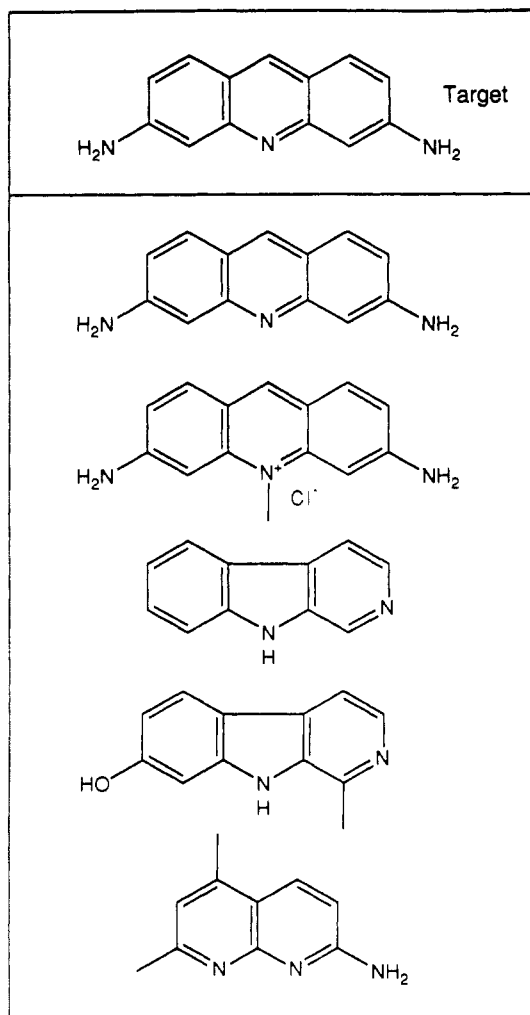


Figure 9. Results of a search for compounds similar to an acridine.

groups in the target structure. Distinguishing between lone pairs on heteroatoms in different environments, as discussed under Enhancements above, would make this possible.

Adenine gave the results shown in Figure 10, which require no comment. The final target structure to be discussed is 3-bromothiophene, which gave the results shown in Figure 11. The listing of structures in the same row in Figure 11 is not intended to imply equal ranking, but has been done to save space. Comments about rankings are made in the following discussion.

The exact match compound was the only 3-halothiophene in the test set. The next-ranked compound was 2-bromothiophene, followed by the chloro and iodo analogues. This illustrates operation of the sorting hierarchy in the system. The primary sort ranked all three compounds equally. The bromo compound was separated out for higher ranking by secondary sorting based on specific halogen type (the chloro and iodo analogues ranked equally with each other).

The sixth compound in the list is 3-bromofuran, marking a departure from rigid adherence to a particular heterocyclic ring. If sulfur and oxygen had been declared as generic types, this compound would have ranked second, ahead of the 2-substituted thiophenes. Two thiazoles appear among the answer set shown in the figure (at ranking positions 8 and 11). They are less similar to the target than the compounds discussed above, but they are valid selections from the pool of widely-differing heterocycles used for the test.

3-Bromopyridine appears in the answer set. This is a valid selection, since it is another aromatic heterocycle substituted

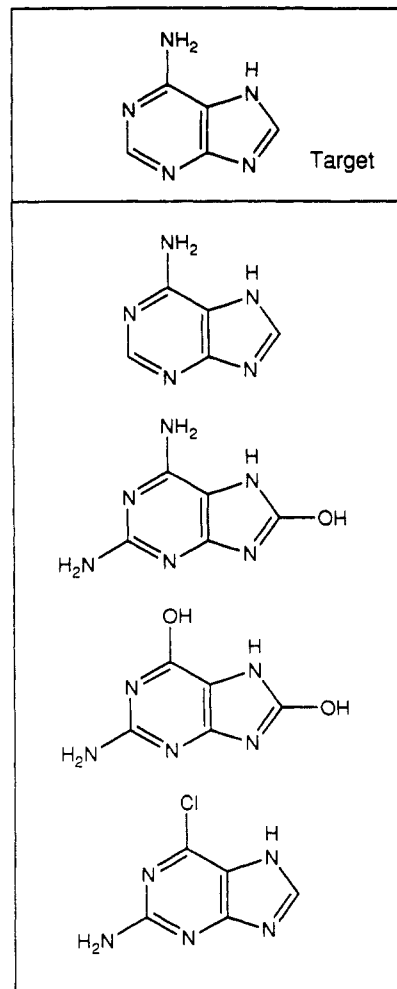


Figure 10. Results of a search for compounds similar to adenosine.

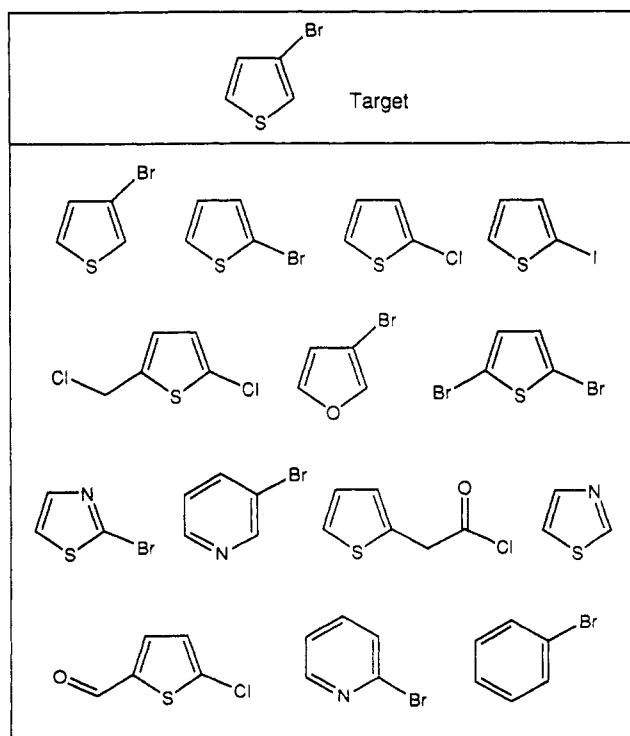


Figure 11. Results of a search for compounds similar to 3-bromothiophene.

by bromine in the 3-position, and it is a selection that a system based on predefined fragments would be unlikely to find.

Bromobenzene also occurs, at a suitably lower ranking. Like the target, it is an aromatic bromo compound. Other similarity methods might not be able to select it in preference to the remaining 168 mainly heterocyclic compounds in the test set.

ACKNOWLEDGMENT

I thank C. Marshall for useful discussions during the course of the work described in this paper.

REFERENCES AND NOTES

- (1) Bawden, D. Computerized Chemical Structure Handling Techniques in Structure-Activity Studies and Molecular Property Prediction. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 14-22.
- (2) MACCS is supplied by Molecular Design Ltd., 2132 Farallon Dr., San Leandro, CA 94577.
- (3) Willett, P.; Wintermann, V.; Bawden, D. Implementation of Nearest-Neighbor Searching in an Online Chemical Structure Search System. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 36-41.
- (4) Enslein, K.; Borgstedt, H. H.; Blake, B. W.; Hart, J. B. Estimation of Maximum Tolerated Dose for Long-Term Bioassays from Acute Lethal Dose and Structure by QSAR. *Risk Anal.* **1991**, *11*, 509-517.
- (5) Klopman, G. Predicting Toxicity through a Computer Automated Structure Evaluation Program. *EHP, Environ Health Persp.* **1985**, *61*, 269-274.
- (6) Judson, P. N. QSAR and Expert Systems in the Prediction of Biological Activity. *Pestic. Sci.*, in press.
- (7) Sheridan, R. P.; Venkataraghavan, R. New Methods in Computer-Aided Drug Design. *Acc. Chem. Res.* **1987**, *20*, 322-329.
- (8) Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological Torsion: A New Molecular Descriptor for SAR Applications. Comparison with Other Descriptors. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 82-85.
- (9) Carhardt, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64-73.
- (10) Adamson, G. W.; Bush, J. A. A Comparison of the Performance of Some Similarity and Dissimilarity Measures in Automatic Classification of Chemical Structures. *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 55-58.
- (11) *Fine Chemicals Directory* is supplied by Molecular Design Ltd., 2132 Farallon Dr., San Leandro, CA 94577.