APPLICATIONS OF INFORMATION RETRIEVAL METHODS IN COMPUTER-AIDED DRUG DISCOVERY

Peter Willett
Krebs Institute for Biomolecular Research and
Department of Information Studies,
University of Sheffield, Western Bank, Sheffield S10 2TN, UK.
Email p.willett@sheffield.ac.uk

ABSTRACT

Measures of similarity play an important role in modern systems for textual information retrieval. This paper reviews the use of such measures for processing databases of chemical structures, which are of increasing importance in the discovery of new pharmaceuticals. Methods are described for chemical similarity searching, for clustering databases of chemical substances, and for selecting structurally diverse database subsets.

Keywords: Chemical databases, Chemical similarity searching, Cluster analysis, Document clustering, Information retrieval, Molecular diversity analysis, Molecular similarity, Search engine

INTRODUCTION

The calculation of similarity lies at the heart of many of the tools comprising modern systems for text-based information retrieval. Most obviously, the ranked-output retrieval facilities in current Web search engines are based on calculating the similarity between a user's query and each of the Web pages that have been indexed by an engine's spider system. The similarity here is based on the number of words in common to the query and page texts, using weighting and normalisation methods that have now been studied for over three decades (see, e.g., Sparck Jones, 2000; Spark Jones and Willett, 1997; van Rijsbergen, 1979). Similar measures are used to calculate the inter-document similarities that lie at the heart of methods for automatic document classification (Jardine and van Rijsbergen, 1971; Salton, 1989; Willett, 1988).

Like many departments of librarianship and information science, the Department of Information Studies at the University of Sheffield has had a long-standing interest in the development of methods for processing textual databases (see, e.g., Lynch and Willett, 1987). It has additionally pioneered a range of techniques for processing the databases of chemical structures that play an increasingly important role in the computer-aided discovery of novel drugs. The Department's dual research focus has led to the realisation that it is often, though by no means invariably, the case that algorithms and data structures that can be applied to one type of database processing can also be applied to the other. There are several reasons for this. Firstly, there are clear similarities in the ways that chemical and textual database records are characterised. The documents in a text database are each typically indexed by some small number of keywords, in just the same way as the molecules in a chemical database are each characterised by some small number of substructural features chosen from a much larger number of potential attributes (as discussed further in the next section of this paper). Moreover, both types of attribute follow a well-marked Zipfian distribution, with the skewed distributions that characterise the frequencies of occurrence of characters, character substrings and words in text databases being mirrored by the comparable distributions for the frequencies of chemical substructural moieties. These shared characteristics mean that the two types of database are amenable to efficient processing using the same types of file structure. Finally, in just the same way as a document either is, or is not, relevant to some particular user query, so a molecule is active, or is not active, in some particular biological test, thus allowing comparable performance measures to be used to assess search effectiveness in chemical and textual retrieval systems.

In a previous paper (Willett, 1999), we have given several examples of the close relationships that exist between textual and chemical information processing. Here, we focus on chemical applications of similarity and clustering methods that were first developed for applications in textual information retrieval. Indeed, it was our experience of these methods in the textual domain that led us to consider their application to chemical searching and clustering (Willett, 1987) and, subsequently, to molecular diversity analysis (Martin *et al*., 2001), with many of these Sheffield, text-derived methods now incorporated in commercial software for chemical information management.

CHEMICAL DATABASE SYSTEMS

Many different scientific disciplines (such as synthetic organic chemistry, structural biology, pharmacology and toxicology) are needed to discover the new drugs that are the lifeblood of the pharmaceutical industry (see, e.g., Landau *et al.* 1999). The huge costs and extended timescales that characterise the industry mean that it is willing and able to make very substantial investments in any technology that can increase the speed with which drugs, *i.e.*, novel chemical molecules with beneficial biological properties, are brought to the market place (and similar comments apply to the pesticides and fungicides developed by the agrochemicals industry). Such investments have provided one of the principal driving forces for the highly sophisticated systems that have been developed for the storage, retrieval and processing of two-dimensional (2D) and three-dimensional (3D) structures of chemical compounds (Ash *et al.*, 1991; Martin and Willett, 1998).

The principal method of representation for a 2D chemical structure diagram is a labelled graph in which the nodes and edges of a graph represent the atoms and bonds, respectively, of a molecule. A chemical database can hence be represented by a large number of such graphs, with searching historically being carried out using two types of graph isomorphism algorithms. *Structure searching* involves an exact-match search of a chemical database for a specific query structure as is required, for example, to retrieve the biological assay results and the synthetic details associated with a particular molecule. Such a search is effected by means of a graph isomorphism search, in which the graph describing the query molecule is checked for isomorphism with the graphs of each of the database molecules. *Substructure searching* involves a partial-match search of a chemical database to find all those molecules that contain a user-defined query substructure, irrespective of the environment in which that substructure occurs. For example, Table 2 shows typical substructure search output, with all of the retrieved molecules containing the diphenyl ether query moiety.

A substructure search is effected by checking the graph describing the query substructure for subgraph isomorphism with the graphs of each of the database molecules (Barnard, 1993). However, subgraph isomorphism is known to be NP-complete and substructure searching in databases of non-trivial size would hence be totally infeasible if it were not for the use of an initial *screen search*, where a screen is a substructural feature, the presence of which is necessary, but not sufficient, for a molecule to contain the query substructure. These features are typically small, atom-, bond- or ring-centred fragment substructures that are algorithmically generated from a connection table when a molecule is added to the database that is to be searched. The fragments that have been chosen for use in screening are listed in a fragment coding dictionary, which will typically contain a few hundred or a few thousand carefully selected fragments (Barnard, 1993). Each of the database structures is analysed to identify those screens from the coding dictionary that are present, and then represented for search by a fixed-length bit-string in which the non-zero bits correspond to the screens that are present. The query (sub)structure is subjected to the same process and the screen search then involves checking the bit-strings representing each database structure for the presence of the screens that are encoded in the bit-string representing the query substructure.

Only a very small fraction of a database will normally contain all of the screens that have been assigned to a query substructure, and thus only these few molecules need to undergo the final, time-consuming search to ensure that there is an exact subgraph isomorphism between the graphs representing the query substructure and each database structure. This simple, two-stage procedure (*i.e.*, screen searching and subgraph searching) has formed the basis for most operational 2D substructure searching systems, and similar techniques are used for 3D substructure searching (Martin and Willett, 1998). The idea of a split-level search is analogous to that used in signature-based systems for serial text scanning, where an initial bit-string search is used to eliminate most of the documents in a database from a time-consuming pattern matching search (see, e.g., Faloutsos, 1985).

CHEMICAL SIMILARITY SEARCHING

Substructure searching, whether in 2D or in 3D, provides an invaluable tool for accessing databases of chemical structures. It does, however, have several limitations that are inherent in the retrieval criterion that is being used, which is that a database structure must contain the *entire* query substructure in precisely the form that has been specified by the user. Firstly, and most importantly, a substructure search requires that the user who is posing the query must already have acquired a well-defined view of what sorts of structures are expected to be retrieved from the database. This is clearly very difficult at the start of an investigation, when perhaps only one or two active structures have been identified and when it is not at all clear which particular feature(s) within them are responsible for the observed activity. Secondly, there is very little control over the size of the output that is produced by a particular query substructure. Accordingly, the specification of a common ring system, such as the benzodiazepine system that forms the nucleus of many tranquillisers, can result in the retrieval of many thousands of compounds from a chemical database. Finally, a substructure search results in a simple partition of the database into two discrete sub-sets (i.e., those structures that contain the query and those that do not) and there is no direct mechanism by which the retrieved molecules can be ranked in order of decreasing probability of activity.

These limitations are entirely analogous to those suffered by Boolean methods for text retrieval (Salton, 1989; Sparck Jones and Willett, 1997). In just the same way as Boolean retrieval has increasingly been complemented, or even supplanted, by best-match retrieval methods in text search engines, so substructure searching has now been augmented by chemical *similarity searching*. Similarity searching requires the specification of an entire target structure, rather than the partial structure that is required for substructure searching. The target molecule is characterised by a set of structural features, and this set is compared with the corresponding sets of features for each of the database structures. Each such comparison enables the calculation of a measure of similarity between the target structure and a database structure, and the database molecules are then sorted into order of decreasing similarity with the target. The output from the search is a ranked list, where the structures that the system judges to be most similar to the target structure are located at the top of the list. Accordingly, if an appropriate measure of similarity has been used, the first database structures inspected will be those that have the greatest probability of being of interest to the user (Carhart *et al*., 1985).

At the heart of any similarity searching system is the measure that is used to quantify the degree of structural resemblance between the target structure and each of the structures in the database that is to be searched. Willett *et al*. (1998) provide an extended review of inter-molecular structural similarity measures, focusing on those that are sufficiently rapid for similarity searching in databases of non-trivial size. The most common measures of this type are based on comparing the fragment bit-strings that are used for 2D substructure searching, so that two molecules are judged as being similar if they have a large number of bits, and hence substructural fragments, in common. A normalised association coefficient, typically the Tanimoto coefficient, is used to give similarity values in the range of zero (no bits in common) to unity (all bits the same) (Willett *et al*., 1998). Specifically, if two molecules have A and B bits set in their fragment bit-strings, with C of these in common, then the Tanimoto coefficient is defined to be

$$\frac{C}{A+B-C}$$

Such a similarity measure is clearly very similar to those employed in text retrieval systems, where coefficients such as the Dice and Cosine coefficients are used to quantify the numbers of words or index terms common to a query and to a database document (Salton, 1989).

While fragment-based measures such as the Tanimoto one above clearly provide a simple, indeed simplistic, picture of the similarity relationships between pairs of structures, they are both efficient (since they involve just the application of logical operations to pairs of bit-strings) and effective (since they have been shown to be capable of bringing together molecules that are judged by chemists to be structurally similar to each other) in operation. The latter characteristic is most surprising, given that the fragments that are used for the calculation of the similarities were originally designed to maximise the efficiency of substructure searching, not the effectiveness of similarity searching. Examples of some of the top-ranked molecules retrieved in a Tanimoto-based 2D similarity search are shown in Figure 2, where it will be seen that the search has been successful in retrieving molecules that are closely related to the target structure; however, there is no single, unifying, common substructure as is the case in a substructure search such as that shown in Figure 1.

Many other types of similarity measure have been described (Dean, 1994; Johnson and Maggiora, 1990), and at least some have been used for database searching; however, none of these measures has proved to be anywhere near as popular as the simple, fragment-based measures described above, and this type of measure is hence assumed in the remainder of this paper unless stated otherwise.

CLUSTERING OF CHEMICAL DATABASES

Random screening has long played an important rôle in lead-discovery programmes. Here, compounds are selected from a database and then tested in a bioassay that determines whether the selected compounds have the biological activity of interest. The identification of an active compound is used to initiate an iterative process in which a similarity search is used to identify structurally related molecules that are tested, in their turn, for activity. Once several such actives have been identified, a query can be defined to enable substructure searching to be carried out.

Considerations of cost-effectiveness dictate that the compounds selected for biological testing in the initial stages of lead-discovery programmes cover the full range of structural types that are available to an organisation, and there has thus been much interest in computer-based methods that can be used to maximise the coverage of structural space. Cluster analysis, or automatic classification, was the first such technique to be used for this purpose.

Cluster analysis is the process of subdividing a group of objects (chemical molecules in the present context) into groups, or clusters, of objects that exhibit a high degree of both intra-cluster similarity and inter-cluster dissimilarity (Arabie *et al.*, 1996; Everitt, 1993). The clustering of document databases so as to identify clusters that contain large numbers of relevant documents has been studied for many years (Jardine and van Rijsbergen, 1971, Willett, 1988) and the analogies between textual and chemical databases noted previously led us to commence an extended programme of research to determine whether comparable methods could be used to cluster chemical databases. The principal aim of the work was to obtain an overview of the range of structural types present within a dataset by selecting one (or some small number) of the molecules from each of the clusters resulting from the application of an appropriate clustering method to that dataset. The representative molecule for each cluster is either selected at random or selected as being the closest to the cluster centroid. These selected compounds are then tested in the bioassay of interest. If a compound proves active it is then appropriate to assay the other compounds in its cluster since these may also exhibit the activity of interest: the fact that structurally similar molecules have similar properties is normally referred to as the *similar property principle* (Johnson and Maggiora, 1990), which is analogous in many ways to the *cluster hypothesis* that provides the principal rationale for the use of clustering in document retrieval (Jardine and van Rijsbergen, 1971).

Very many different clustering methods have been described in the literature, and it was hence necessary to compare the effectiveness of the various methods for clustering chemical structures, typically represented by the fragment bit-strings that are used for substructure and similarity searching as discussed previously. These comparisons used a 'leave-one-out' experimental methodology that was first suggested by Adamson and Bush (1973) and that is based upon the similar property principle. Assume that the value of some quantitative (*i.e.*, interval or ratio scale) property has been measured for each of the molecules in a dataset. The property value of a molecule, *I*, within this dataset is assumed to be unknown, and the classification resulting from the use of some particular clustering method is scanned to identify the cluster that contains the molecule *I*. The predicted property value for *I*, *P*(*I*), is then set equal to the arithmetic mean of the observed property values of the other compounds in that cluster. This procedure results in the calculation of a *P*(*I*) value for each of the *N* structures in the dataset, and an overall figure of merit for the classification is then obtained by calculating the product moment correlation coefficient between the sets of *N* observed and *N* predicted values. The most generally useful clustering methods will be those that give high correlation coefficients across as wide a range of datasets as possible.

Adamson and Bush's approach to the comparison of clustering methods was used by Willett (1987) in a study of over 30 hierarchic and non-hierarchic clustering methods when applied to 10 small datasets for which physical, chemical or biological property data were available. The study found that the best results were obtained with Ward's hierarchic-agglomerative method (Ward, 1963), with the non-hierarchic nearest-neighbour method of Jarvis and Patrick (1973) performing almost as well. At the time that these comparative experiments were carried out, computer limitations (in terms of both raw CPU speeds and the algorithms available) meant that Ward's method could not be applied to chemical databases of substantial size. The Jarvis-Patrick method was thus rapidly adopted as the clustering method of choice in commercial chemical database software, not only to select compounds for random screening but also to cluster the outputs of substructure searches that retrieve very large numbers of molecules, thus providing the searcher with an overview of the structural classes that contain the substructure of interest (Willett, 1987). However, the method does have limitations and subsequent comparisons (Brown and Martin, 1996, 1997; Downs *et al.*, 1994) have reaffirmed the general superiority of Ward's method. The availability of improved computer hardware and of the efficient reciprocal nearest neighbours algorithm (Murtagh, 1985) means that this method can now be applied to databases containing some hundreds of thousands of molecules in an acceptable amount of time, and Ward's method is thus becoming available in commercial chemical database software; larger datasets, however, still require use of the Jarvis-Patrick method.

MOLECULAR DIVERSITY ANALYSIS

Pharmaceutical research has been revolutionised over the last few years by the emergence of *combinatorial chemistry* (see, e.g., DeWitt and Czarnik, 1997), a body of techniques for the parallel synthesis and testing of sets of molecules, called *combinatorial libraries*, that contain large numbers (hundreds or thousands) of structurally related molecules. Such techniques are increasingly replacing the traditional approach to drug discovery, which involved a sequential mode of processing with molecules being synthesised and then tested for biological activity one molecule at a time.

The need to ensure coverage of the largest possible expanse of chemical space in the search for bioactive molecules means that combinatorial approaches seek to maximise the *diversity* of chemical libraries, *i.e.*, the degree of structural variation that is present within the set of product molecules resulting from a combinatorial synthesis, whilst ensuring that as few compounds as possible should be selected for synthesis and biological testing on grounds of cost-effectiveness. The concept of diversity is normally quantified using similarity-based techniques that are a natural development of those discussed above, with a diverse set of molecules being selected by consideration of their Tanimoto-based inter-molecular similarities.

There is a trivial algorithm available to identify the most diverse *n*-compound subset of an *N*-compound database or library (where, typically, $n \ll N$). This algorithm involves generating each of the

$$\frac{N!}{n!(N-n)!}$$

possible subsets and then calculating their diversities using a *diversity index* (some function of the inter-molecular similarities in the chosen subset): the optimal subset is then that group of compounds that has the greatest value of the diversity index. Such a procedure is computationally infeasible for realistic values of *n* and *N* and there has thus been much interest in alternative approaches for selecting diverse sets of molecules (Dean and Lewis, 1999; Ghose and Viswanadhan, 2001). Cluster-based selection, as described in the previous section, was the first such approach to be used (see, e.g., Shemetulskis *et al.*, 1995) but it is now increasingly being complemented by alternative approaches. One such approach is dissimilarity-based compound selection, the basic algorithm for which was first described by Lajiness (1990) and Bawden (1993). The Bawden-Lajiness algorithm involves selecting a compound at random and then iteratively choosing that previously unselected compound that is most dissimilar to those that have already been selected. The algorithm is very simple in concept but has an expected time complexity of order $O(n^2N)$ for selecting an *n*-compound subset from an *N*-compound dataset, and might hence be too time-consuming for large-scale applications.

It was at this point that we turned again to work on hierarchic document clustering. The various hierarchic clustering methods differ only in the precise criterion that is used to measure the similarity between two clusters of documents at each stage in the creation of a hierarchy. The appropriate criterion for the inter-cluster similarity in the group-average method is the average of all of the pairwise inter-document similarities, where one document is in one of the two selected clusters and the other document is in the other cluster. Voorhees (1986) demonstrated that precisely the same average similarity could be obtained from a procedure that involved just a single similarity calculation using the weighted centroids of the two clusters. This elegant equivalence provides a highly efficient way of implementing the group-average method; however, we realised that it can also be applied much more generally to any situation where: sums of similarities, rather than individual similarities, are required; where the cosine coefficient is used to measure the similarity between pairs of objects; and where the objects that are being compared are characterised by some form of vector-like representation (such as fragment bit-strings). Specifically, Voorhees' result can be used to choose the most dissimilar molecule in the selection step of the Bawden-Lajiness algorithm, if by "most dissimilar" we mean that molecule with the largest sum of dissimilarities to the molecules that have already been chosen). This results in an algorithm with an expected time complexity of $O(nN)$, thus allowing the Bawden-Lajiness algorithm to be used on very large files of compounds (Holliday *et al*., 1995). It should be noted that later studies have suggested the superiority of alternative dissimilarity-based selection algorithms (Agrafiotis and Lobanov, 1999; Snarey *et al*., 1998); however, our work does provide yet another example of the ways in which approaches first designed for processing text databases are applicable to the very different domain of computer-aided drug discovery.

Acknowledgements. I thank all of my colleagues, past and present, for their contributions to the research that has been summarised here, and the many organisations that have provided financial support for my research over the years. The Krebs Institute for Biomolecular Research is a designated Biomolecular Sciences Centre of the Biotechnology and Biological Sciences Research Council.

REFERENCES

Adamson, G.W. and Bush, J.A. (1973) A method for the automatic classification of chemical structures. *Information Storage and Retrieval*, **9**, 561-568.

Agrafiotis, D. and Lobanov, V.S. (1999) An efficient implementation of distance-based diversity measures based on *k-d* trees. *Journal of Chemical Information and Computer Sciences*, **39**, 51-58.

Arabie, P., Hubert, L.J. and De Soete, G. (1996) (editors) *Clustering and Classification*. Singapore: World Scientific.

Ash, J.E., Warr, W.A. and Willett, P. (1991) (editors) *Chemical Structure Systems*. Chichester: Ellis Horwood.

Barnard, J.M. (1993) Substructure searching methods: old and new. *Journal of Chemical Information and Computer Sciences*, **33**, 532-538.

Bawden, D. (1993) Molecular dissimilarity in chemical information systems. In: Warr, W.A. (ediror) *Chemical Structures 2. The International Language of Chemistry*. Heidelberg: Springer-Verlag.

Brown, R.D. and Martin, Y.C. (1996) Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. *Journal of Chemical Information and Computer Sciences*, **36**, 572-584.

Brown, R.D. and Martin, Y.C. (1997) The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding. *Journal of Chemical Information and Computer Sciences*, **37**, 1-9.

Carhart, R.E., Smith, D.H., Venkataraghavan, R. (1985) Atom pairs as molecular features in structure-activity studies: definition and applications. *Journal of Chemical Information and Computer Sciences*, **25**, 64-73.

Dean, P.M. (1994) (editor) *Molecular Similarity in Drug Design*. Glasgow: Chapman and Hall, 1994.

Dean, P.M. and Lewis, R.A. (1999) (editors) *Molecular Diversity in Drug Design*. Amsterdam: Kluwer

DeWitt, S.H. and Czarnik, A.W. (1997) (editors) *A Practical Guide to Combinatorial Chemistry*. Washington DC: American Chemical Society.

Downs, G.M., Willett, P. and Fisanick, W. (1994) Similarity searching and clustering of chemical-structure databases using molecular property data. *Journal of Chemical Information and Computer Sciences*, **34**, 1094-1102.

Everitt, B.S. (1993) *Cluster Analysis*. 3rd edition. London: Edward Arnold.

Faloutsos, C. (1985) Access methods for text. *Computing Surveys*, **17**, 49-74.

Ghose A.K. and Viswanadhan V.N. (2001) (editors) *Combinatorial Library Design and Evaluation for Drug Discovery: Principles, Methods, Software* Tools and Applications. New York: Marcel Dekker.

Holliday, J.D., Ranade, S.S. and Willett, P. (1995) A fast algorithm for selecting sets of dissimilar structures from large chemical databases. *Quantitative Structure-Activity Relationships*, **14**, 501-506.

Jardine, N. and van Rijsbergen, C.J. (1971) The use of hierarchic clustering in information retrieval. *Information Storage and Retrieval*, **7**, 217-240.

Jarvis, R.A. and Patrick, E.A. (1973) Clustering using a similarity measure based on shared nearest neighbours. *IEEE Transactions on Computers*, **C-22**, 1025-1034.

Johnson, M.A. and Maggiora, G.M. (1990) (editors) *Concepts and Applications of Molecular Similarity*. New York: Wiley.

Lajiness, M. (1990) Molecular similarity-based methods for selecting compounds for screening. In: Rouvray, D.H. (editor) *Computational Chemical Graph Theory*. New York: Nova Science Publishers.

Landau, R., Achilladelis, B. and Scriabine, A. (1999) (editors) *Pharmaceutical Innovation*. Philadelphia PA: Chemical Heritage Foundation.

Lynch, M.F. and Willett, P. (1987) Information retrieval research in the Department of Information Studies, University of Sheffield: 1965-1985. *Journal of Information Science*, **13**, 221-234.

Martin. Y.C. and Willett, P. (1998) (editors) *Designing Bioactive Molecules: Three-Dimensional Techniques and Applications*. Washington: American Chemical Society.

Martin, Y.C., Willett, P., Lajiness, M., Johnson, M., Maggiora, G., Martin, Y.C., Bures, M.G., Gasteiger, J., Cramer, R.D., Pearlman, R.S. and Mason, J.S. (2001) Diverse viewpoints on computational aspects of molecular diversity. *Journal of Combinatorial Chemistry*, **3**, 231-250.

Murtagh, F. (1985) *Multidimensional Clustering Algorithms*. Vienna: Physica Verlag.

Salton, G. (1989) *Automatic Text Processing*. Reading, MA: Addison-Wesley.

Shemetulskis, N.E., Dunbar, J.B., Dunbar, B.W., Moreland, D.W. and Humblet, C. (1995) Enhancing the diversity of a corporate database using chemical database clustering and analysis. *Journal of Computer-Aided Molecular Design*, **9**, 407-416.

Snarey, M., Terret, N.K., Willett, P. and Wilton, D.J. (1998) Comparison of algorithms for dissimilarity-based compound selection. *Journal of Molecular Graphics and Modelling*, **15**, 372-385.

Sparck Jones, K. (2000) Further reflections on TREC. *Information Processing and Management*, **36**, 37-85.

Sparck Jones, K. and Willett, P. (1997) (editors) *Readings in Information Retrieval*. San Francisco: Morgan Kaufmann.

van Rijsbergen, C.J. (1979) *Information Retrieval*. London: Butterworth

Voorhees, E.M. (1986) Implementing agglomerative hierarchic clustering algorithms for use in document retrieval. *Information Processing and Management*, **22**, 465-476.

Ward, J.H. (1963) Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, **58**, 236-244.

Willett, P. (1987) *Similarity and Clustering in Chemical Information Systems*. Letchworth, Research Studies Press.

Willett, P. (1988) Recent trends in hierarchic document clustering: a critical review. *Information Processing and Management*, **24**, 577-597.

Willett, P. (1999) Textual and chemical information retrieval: different applications but similar algorithms. *Information Research*, **5**(2), 1999, at URL http://InformationR.net/ir/5-2/infres52.html

Willett, P., Barnard, J.M. and Downs, G.M. (1998) Chemical similarity searching. *Journal of Chemical Information and Computer Sciences*, **38**, 983-996.

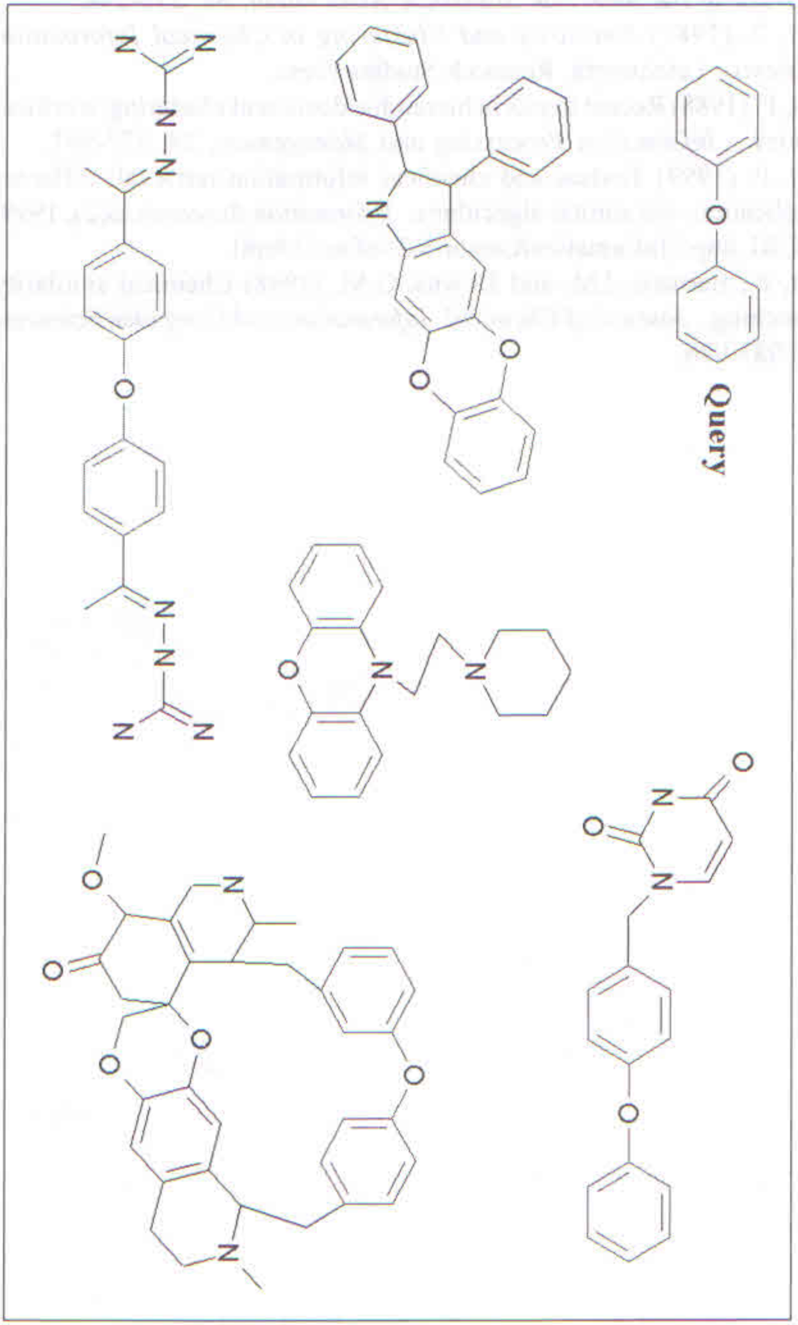

Figure 1: Sample output from a 2D substructure search

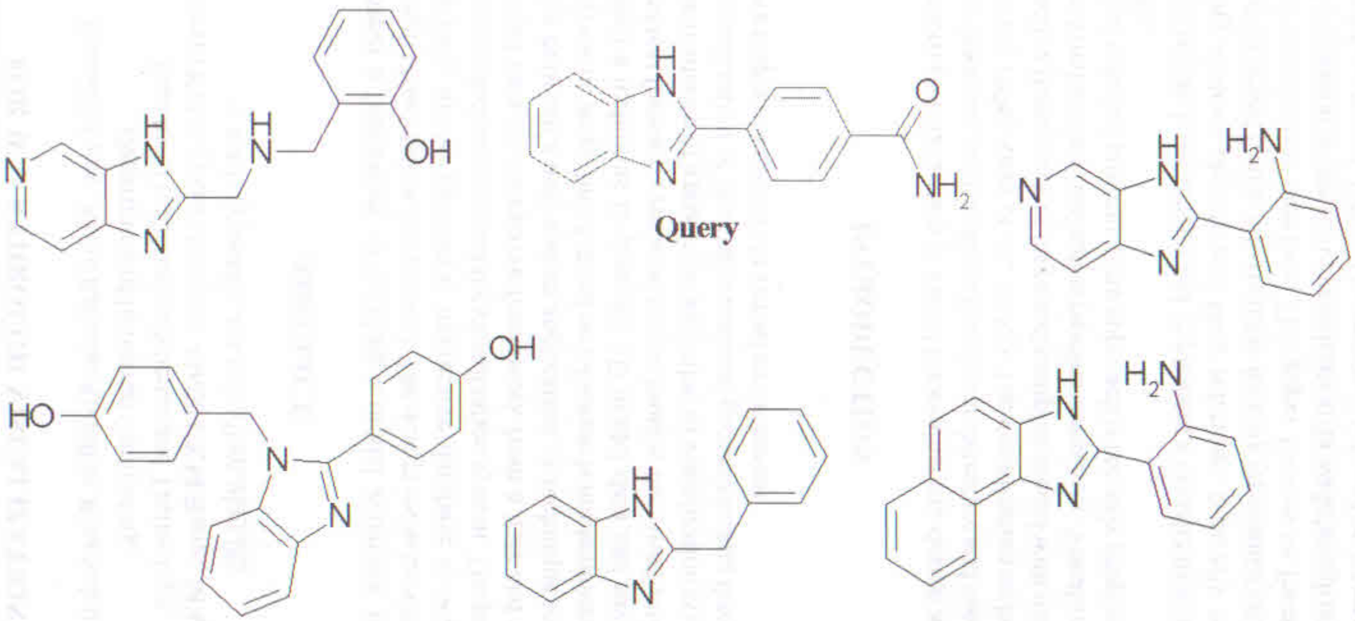

Figure 2: Sample output from a 2D similarity search