
Article:
Bishop, N., Gillet, V.J., Holliday, J.D. et al. (1 more author) (2003) Chemoinformatics Research at the University of Sheffield: A History and Citation Analysis. Journal of Information Science, 29 (4). pp. 249-267. ISSN 0165-5515

https://doi.org/10.1177/01655515030294003

Chemoinformatics Research at the University of Sheffield: A History And Citation Analysis

Neal Bishop, Valerie J. Gillet,

John D. Holliday and Peter Willett[1]

Department of Information Studies, University of Sheffield,

Western Bank, Sheffield S10 2TN, UK

Abstract. This paper reviews the work of the Chemoinformatics Research Group in the Department of Information Studies at the University of Sheffield, focussing particularly on the work carried out in the period 1985-2002. Four major research areas are discussed, these involving the development of methods for: substructure searching in databases of three-dimensional structures, including both rigid and flexible molecules; the representation and searching of the Markush structures that occur in chemical patents; similarity searching in databases of both two-dimensional and three-dimensional structures; and compound selection and the design of combinatorial libraries. An analysis of citations to 321 publications from the Group shows that it attracted a total of 3725 residual citations during the period 1980-2002. These citations appeared in 411 different journals, and involved 910 different citing organisations from 54 different countries, thus demonstrating the widespread impact of the Group's work.

Keywords. Bibliometrics, Bioinformatics, Chemical database, Chemical patent, Chemical structure, Citation analysis, Cluster analysis, Combinatorial library design, Drug discovery, Genetic algorithms, Graph theory, Markush structure, Molecular diversity analysis, Pharmacophore searching, Similarity searching, Substructure searching

[1] To whom all correspondence should be addressed. Email: p.willett@sheffield.ac.uk

1. Introduction

Chemoinformatics is the name that is increasingly being given to the computational methods that are used to support the discovery of novel biologically active chemical molecules, most notably pharmaceuticals and agrochemicals; other names for this specialism include cheminformatics, chemical information management and chemical structure handling. Although the name is quite new, the subject is one that has existed for very many years [1], with the first practicable system for substructure searching (*vide infra*) being described by Ray and Kirsch as early as 1957 [2]. The Department of Information Studies in the University of Sheffield has been working in this area for almost all of the forty years that the Department has been in existence and that are celebrated in this special issue of the *Journal of Information Science*. This long-standing interest arose from the appointment to the Department in 1965 of Michael F. Lynch, following a term as the Head of Basic Research at Chemical Abstracts Service (CAS) at a time when CAS was carrying out some of the first experiments anywhere in the world on the use of computers for the processing of both textual and chemical structural information (see, e.g., [3, 4]). On arrival in Sheffield, Lynch rapidly established research programmes in both of these areas (see, e.g., [5, 6]) and this dual focus of computational activity has continued with great success right up to the present day, with research in one of the areas often providing the spur for subsequent developments in the other [7].

Early chemoinformatics research, in the period 1965 to 1985, focused on the development of methods for indexing databases of chemical reactions, for the design of two-dimensional (2D) substructure searching systems and for the correlation of substructural occurrence data with physical and chemical properties. This early work is reviewed by Lynch and Willett [8] and the present paper hence focuses on research subsequent to that period. The next section describes our studies of substructure searching on databases of three-dimensional (3D) chemical structures, and Section 3 then discusses the techniques that were developed for representing and searching the generic chemical structures that occur in chemical patents, These areas highlight the long-term nature of much of our research since the two sets of studies reported here both extended

over more than a decade. Further studies of database retrieval are summarised in Section 4, with much of the work on similarity searching described in this section providing a natural starting point for the development of a range of techniques for compound selection and for the design of combinatorial libraries, as summarised in Section 5. In much the same way, our earlier studies of 3D substructure searching have resulted in the development of a range of bioinformatics methods for the representation and searching of biological macromolecules (as recently reviewed by Artymiuk *et al*. [9]). The citation analysis in Section 6 demonstrates the extent to which the work in Sheffield has been taken up by other workers, both academic and industrial, and the paper concludes with a brief summary of our current research activities.

2. 3D substructure searching

Substructure searching involves searching a database of chemical structures to identify those that contain some user-defined query pattern. For example, Figure 1 shows five of the molecules that were retrieved in a search of 117,659 molecules from the National Cancer Institute (NCI) database (at URL http://dtp.nci.nih.gov/); this search, in which each of the constituent atoms of the query substructure could be substituted by any number of additional atoms, resulted in the retrieval of a total of 2,737 molecules containing the diphenyl ether query substructure shown at the top of the figure.

Figure 1 about here

Methods for searching databases of 2D chemical structure diagrams are based on the use of graph-theoretic methods, in which the nodes and edges of a graph are used to denote the atoms and bonds of a molecule [10]. Substructure searches can then be effected by means of a *subgraph isomorphism* algorithm, which determines whether the graph denoting the query substructure is a subgraph of the graphs representing each of the database structures. This provides a highly effective way of identifying query substructures but is extremely demanding of computational resources, as subgraph isomorphism belongs to the class of computational problems that are known to be NP-complete. Efficient searching is achieved by means 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, the presence of which characterise a molecule in much the same way that an index term characterises a textual document. The incidences of these fragments in a molecule are encoded in a bit-string (or *fingerprint*) that can be searched extremely rapidly, and that thus eliminates from further consideration large numbers of molecules that cannot possibly satisfy the subgraph match.

Gund [11] was the first to suggest that such graph-based search methods could also be applied to the retrieval of 3D chemical structures, with the nodes and edges of a graph being used to represent the atoms and inter-atomic distances, respectively, in a 3D molecule. The resulting inter-atomic distance matrix could then be inspected for the presence of a query *pharmacophore*, or *pharmacophoric pattern*, *i.e*., the arrangement of structural features in 3D space necessary for a molecule to bind at an active site; an example of an anti-leukemic pharmacophore [12] is shown in Figure 2. Given that a 3D structure can be represented by a graph, the presence or absence of a pharmacophoric pattern can be confirmed by means of a subgraph isomorphism procedure in which the edges in a database structure and a query substructure are matched if they denote the same inter-atomic distance (to within any user-specified tolerance such as ±0.5Å).

Insert Figure 2 about here

Our initial studies of 3D substructure searching focused on the design of a screening system to permit rapid searching of large databases [13, 14]. Given that pharmacophores are normally expressed in terms of inter-atomic distances, the screens that we developed consisted of a pair of atoms together with an associated inter-atomic distance range. For example, a typical screen might consist of an oxygen atom separated from a nitrogen atom by a distance of between 6.5 and 8.2 Å: any molecule containing these two atom-types separated by a distance within this range would be assigned that screen, while if the distance was outside the range it would be assigned one of the other oxygen-nitrogen screens. Similar ideas have been adopted by subsequent workers [15, 16]. Once the initial screen search has been carried out, those molecules that contain the screens associated with the query pharmacophore are passed on for the subgraph isomorphism

search [17, 18]. Studies of several different subgraph isomorphism algorithms [19] demonstrated the general efficiency of that described by Ullmann [20] for chemical applications; this algorithm now forms the basis for many operational substructure searching systems, both 2D and 3D.

Although both effective and efficient in operation, these screening and geometric searching algorithms are limited in that they take no account of the flexibility that characterises many molecules [18, 21]. Specifically, early 3D searching systems stored only a single low-energy *conformation* (i.e., a particular arrangement of the atoms in a molecule in 3D space) for each molecule in a database, with the result that a pharmacophore search is likely to miss large numbers of matching molecules that can adopt a conformation containing the query pattern but that are represented in the database by a low-energy conformation that does not contain this pattern. Two main approaches to flexible 3D searching have been described in the literature [21]. In the first, a flexible molecule is represented by some small number of carefully selected low-energy conformations, with a rigid-searching algorithm being applied to each of the conformations describing a molecule. This approach has many advantages but it does mean that the search algorithm cannot explore the full conformational space available to a flexible molecule, with the possibility of a loss in recall. The work in Sheffield thus focused upon the development of algorithms and data structures that could avoid such retrieval failures.

In a rigid 3D molecule, the distance between each and every pair of atoms is a single, fixed value, whereas the distance between a pair of atoms in a flexible molecule will depend on the conformation that is adopted. The separation of a pair of atoms is hence conveniently described by a *distance range*, the lowerbounds and upperbounds of which correspond to the minimum- and maximum-possible distances. The set of distance ranges for a molecule will contain all of the geometrically-feasible conformations which that molecule can adopt, and thus provides an obvious way of representing a flexible molecule. Such sets of distance ranges can be generated using the bounds-smoothing

technique that forms an important component of the *distance-geometry* approach to structure generation [22].

The screening and graph-searching algorithms that are used for rigid 3D searching operate on graphs where each edge denotes a single value; these algorithms require only minor modifications to enable them to process graphs in which each edge contains both a lowerbound and an upperbound, thus allowing the retrieval of all molecules that could possibly adopt a conformation that contains a query pharmacophoric pattern [23]. Indeed, it is possible to view the algorithms described previously for rigid searching as limiting cases of the more general algorithms that are required for flexible searching. There is, however, one major difference between flexible 3D and both 2D and rigid 3D substructure searching, in that those molecules that match the query in the subgraph-isomorphism search must then undergo a further, and final, check that uses some form of conformational-searching procedure; this is required since bounds-smoothing is known to over-estimate the true range of possible inter-atomic distances. A range of methods for this final conformational search have been described [24], of which the most effective and most efficient would seem to be a technique known as *directed tweak* [25].

Flexible 3D pharmacophore searching is now well-established and plays an important role in lead-discovery programmes for novel pharmaceutical and agrochemical compounds (see, e.g., [26, 27]). As an example, Figure 3 shows some of the hits from a 3D search for the anti-leukemic pharmacophore of Figure 2 against the NCI file mentioned previously. This search retrieved a total of 1341 hits when the molecules were represented by just a single conformation; five of these hits are shown in Figure 3, where it will be seen that they encompass a wide range of structural types containing the specified three atoms at distances within the allowed tolerances (as marked by the dotted lines). The number of hits increased to no less than 5541 when the corresponding flexible search was carried out (using the UNITY system for chemical information management [28]), thus illustrating the much greater level of recall that can be achieved if account is taken of conformational flexibility.

Insert Figure 3 about here

As well as providing the basis for current operational systems for 3D substructure searching, the research described above also provided the principal initial stimulus for an extended collaboration with the Department of Molecular Biology and Biotechnology in the University of Sheffield. This work has developed graph-theoretic methods for the representation and searching of biological macromolecular structures, thus providing a way of analysing biological structures, rather than the biological sequences that lie at the heart of most bioinformatics research. Most of our work in this area has focused on the representation and searching of the 3D protein structures in the Protein Data Bank (at URL http://www.rcsb.org/pdb) but we have also studied carbohydrate and RNA structures; we will not discuss this work further here as a recent overview is provided by Artymiuk *et al*. [9].

3. Searching generic chemical structures

Patents form one of the most important sources of chemical information. For many years, however, computer-based access to structural information in the patent literature had to be based on fragmentation codes, in which the structural information was characterised by a series of substructural fragments, manually encoded by skilled patent analysts. A 1978 report by the British Library Research and Development Department [29] highlighted the need for improved means of access to this information, in particular to the *generic*, or *Markush*, structures that occur very frequently in patents and that encode many, or even an infinite number of, different specific molecules in a single representation. A typical generic structure is shown in Figure 4.

Insert Figure 4 about here

There are four principal sources of structural variation that can be encoded in a patent (as discussed by Dethlefsen *et al*. [30, 31]). Substituent variation relates to the variety of possible substituents at a fixed position on a partial structure, e.g., “phenyl substituted in para position by F, Cl or Br”. Position variation relates to the choice of attachment position of a substituent, e.g., “monochlorophenyl”. Frequency variation relates to the frequency of occurrence of substructures, e.g., “CH3-(CH2)n –Cl; n=1-3” (denoting the

ethyl, propyl and n-butyl groups). Homology variation relates to substructures described in terms of chemical families by terms such as “cycloalkyl” or “six-membered heterocycle containing one nitrogen, oxygen or sulphur atom”. Further features are often superimposed on the overall structure, such as nesting in which one substituent may itself contain further structural variation. The combinatorial complexities produced by these variation types results in a document that may describe a large, possibly infinite, number of individual compounds in a single patent claim.

These complexities were the starting point for a programme of research in the Department that started in 1979 and extended over some 15 years. These studies resulted in a body of algorithms and data structures that provided much of the theoretical and practical basis for the sophisticated MARPAT [32, 33] and Markush DARC [34] systems that are the current methods of choice for structure-based access to generic chemical structures. The Sheffield work involved very substantial modifications to the connection-table, screening and atom-by-atom procedures of conventional substructure searching systems: the historical development of the research has been reviewed by Lynch and Holliday [35] so we present here just a brief summary of the major findings.

Lynch and Holliday identified six major scientific achievements of the work. These were: *GENSAL*, a formal and unambiguous language used to describe generic structures for computer input; the *GENSAL Interpreter*, enabling translation of the GENSAL language into an internal computer representation; the *ECTR* (Extended Connection Table Representation), the internal computer representation of the generic structure; the derivation of substructural fragment and ring descriptors used in the initial bit-screening search stage; *reduced chemical graphs*, an intermediate screening stage based on a generalised representation of the structure; and the *refined search*, the final atom-by-atom search component.

GENSAL [36] was designed to take full account of the mechanisms used to describe the structural diversity inherent in chemical patents, and was based on a context-free grammar, similar to many modern day programming languages. The GENSAL

interpreter [37] translates the GENSAL language into the internal representation, the ECTR [38], which is completely unambiguous and encodes all of the logical and structural features described by GENSAL, thus providing an accurate and complete machine-readable description of the specific molecules implied by a generic structure.

The ECTR comprises structural information, positional information, frequency information and logical information in an inverted tree-structured graph representation, the root of which is the invariant part of the molecule. Nodes of the graph describe either the partial structures themselves or their logical relationship to each other. Nodes which represent specific partial structures comprise partial connection tables, while nodes representing homologous series identifiers describe the types of homology present using a list of pre-defined parameters such as the total number of atoms, the number of carbon atoms, etc. Linking these partial structure nodes are the nodes which represent their logical relationship to each other. They may be OR nodes, describing partial structures which are alternatives to each other, or they may be AND nodes, describing partial structures which are in combination. Both types of node also contain much of the variant and invariant inter-component information, such as ranges of frequency or positions of attachment.

The ECTR is the representation from which all subsequent representations, fragment and ring bit-screens and reduced graphs, are derived. The Sheffield screening system used a selection of Augmented Atom, Atom Sequence and Bond Sequence descriptors from the CAS Online dictionary [39] as a test set, together with a set of ring descriptors defined by Downs et al. [40]. The generation of the fragments [41] starts from the terminal nodes of the ECTR and rises up through the tree structure, passing up through every partial structure node to the root node, until all of the constituent fragments have been generated. The fragments need to reflect the environments from which they are generated, particularly as regards whether they are common to all structures described by the generic structure, or whether they are optional. This is facilitated by the ‘bubble-up’ procedure [42], which ensures that the aggregation of features accurately maintains the logical relationships as it moves up the tree.

Fragments are generated from generic partial structures [43] by assigning specific derived parameters to each descriptor in the fragment dictionary, and comparing these with the parameter lists of the respective ECTR representation. The result of the bubble-up is two sets of fragments, one set being common to all specific structures represented by the generic structure, the other being those that are found at least once, but that may be optional. These two sets of fragments are encoded in two separate bit-strings that are searched in much the same way as the bit-string stage of a conventional 2D or 3D substructure search. Subsequent, more detailed searching makes used of reduced chemical graphs [44]. These graphs are produced by fragmenting the structure into distinct components which represent groups of atoms and whose relationship to each other can be maintained: these components are cyclic and acyclic systems, the latter being sub-divided into groups of carbon atoms and groups of non-carbon atoms. These components become nodes in the reduced graph and are further annotated by parameters similar to those used to represent homologous series identifiers in the ECTR, yielding a generalised graph in which the nodes are described using a common representation. The application of a graph matching algorithm to the reduced graphs produced, for each successful query/database mapping, a list of pairs of matching reduced graph nodes. These nodes relate to parts of the ECTR which are represented by real atoms, by parameter lists, or, in some cases, by a mixture of the two. The final, refined search [45] matches the ECTR representations of each pair of nodes at the atom-atom, atom-parameter, or parameter-parameter level as appropriate, using a sophisticated adaptation of the Ullmann subgraph isomorphism algorithm [20].

Following initial funding by the British Library, the work was carried out in collaboration with, and with funding from, groups at Chemical Abstracts Service, Derwent Publications Ltd., and International Documentation in Chemistry (IDC). All three of these groups subsequently implemented operational systems that drew upon the Sheffield work, CAS and Derwent developing the public MARPAT and Markush DARC systems described previously, and IDC developing an in-house database registration system. The studies summarised here thus provide a textbook example of how something that started as

publicly-funded basic academic research can, with appropriate private-sector support, lead to fully functional, widely used, operational information systems.

4. Similarity searching

For many years, substructure searching provided the principal means of access to databases of 2D (and latterly 3D) chemical structures. It does have several inherent limitations and chemical *similarity searching* has arisen as a complementary means of database access that can overcome, or at least alleviate, these limitations (in much the same way as best-match text retrieval has been developed as a way of overcoming some of the inherent limitations of Boolean searching [46, 47]). Thus, a substructure search requires the user to specify precisely the substructural constraints that must be obeyed if a molecule is to be retrieved, and it may accordingly be difficult to define an appropriate query substructure if, *e.g.*, only a single active structure has been identified thus far in a synthetic programme. It is also generally difficult to control the size of the output that is produced, and it is not normally possible to rank the output in order of decreasing utility, even if an output of an appropriate size has been achieved.

Similarity searching involves the user submitting an entire query molecule, normally referred to as the *target structure*, this typically being a molecule that has previously exhibited activity in a biological screening experiment. The similarity search calculates a measure of similarity between the target structure and each of the molecules in the database, and then ranks the database in order of decreasing similarity with the target structure. The Similar Property Principle [48] states that molecules that are structurally similar are likely to have similar properties (in much the same way as the Cluster Hypothesis states that similar documents are likely to be relevant to the same queries [49]). Thus, if a bioactive target structure is searched for, then the top-ranked molecules, which are normally referred to as *nearest neighbours*, are also likely to possess that activity: these molecules are hence prime candidates for biological testing, as compared to other molecules that occur further down the ranking. This *virtual screening* approach provides an attractive way of prioritising the time-consuming and expensive biological

testing that characterises much pharmaceutical and agrochemical research. It does, however, require that the similarity measure that is used be *effective*, i.e., a similarity measure for which high computed structural similarities do, indeed, correspond to similar bioactivity characteristics, and *efficient*, i.e., enable the measure to be calculated sufficiently rapidly for interactive access to large structure databases. The Sheffield group has devoted considerable time during the period under review to the development of similarity measures that exhibit these two, sometimes conflicting, characteristics.

A similarity measure has three main components: the structural *representation* that is used to characterise the molecules that are being compared; the *weighting scheme* that is used to differentiate more important features from less important features; and the *similarity coefficient* that is used to quantify the degree of similarity between pairs of molecules. Of these, the first is probably the most important since the representation that is chosen will control the operations that can be carried out when determining the similarity between a pair of molecules. Much of the early work in Sheffield focused on one particular representation, *viz* the fragment substructures that were first developed for the screening stage of 2D substructure searches. The basic idea underlying the use of this representation is very simple: a database structure is assumed to be similar to a target structure if they have many fragment substructures in common. This idea was first suggested by Adamson and Bush as early as 1973 [50] but only in the context of processing small numbers of molecules, and it was not till the mid-Eighties that it was developed for database searching in two papers that appeared within a few months of each other: one from Lederle Laboratories in the USA [51] and the other from Sheffield [52]. These early studies demonstrated the general utility of this approach to the calculation of inter-molecular structural similarities, and despite the many other measures that have been described since then [48, 53, 54] the combination of fragment substructures and one specific association coefficient, the Tanimoto Coefficient, continues to be the method of choice for similarity searching in operational chemical information systems of all sorts. An example of the nearest neighbours obtained in a Tanimoto-based similarity search is shown in Figure 5, where the close relationship to the target structure is clearly evident.

Figure 5 about here

Similarity searching involves comparing one structure (the target structure) with all of the structures in a database, and it was natural to consider the extension of these ideas to the clustering of chemical structures, where many of the more common methods for cluster analysis involve matching all of the members of a database with each other. Structure-based approaches to the clustering of chemical structures were first suggested in the late Sixties [55, 56] but it was till the Eighties that the Sheffield group began an extended evaluation of the effectiveness of over 30 different types of clustering method when used for the grouping of chemical structures. This work, which is summarised in [57], resulted in the recommendation of the clustering method due to Jarvis and Patrick [58] (when used with fragment substructures and the Tanimoto Coefficient) as being the most suitable for applications such as the selection of compounds for biological testing and the analysis of large substructure search outputs [59], a conclusion that was rapidly taken up in operational chemical information systems (see, e.g., [60]). Later work, both in Sheffield [61] and elsewhere [62], has suggested that Ward's hierarchic agglomerative method [63] may be more effective for chemical clustering, given an efficient algorithm for its implementation.

Later work in Sheffield [64] evaluated the inter-atomic distance screens that are used for 3D substructure searching when applied to the calculation of 3D structural similarities, and this occasioned several subsequent discussions of the use of both distance-based and angle-based methods for 3D similarity searching (see, e.g., [65-67]). However, most of the research, both in Sheffield and elsewhere, of 3D similarity measures has focused on two alternative approaches: the use of the maximum common substructure (MCS) and of molecular field overlaps. Both of these are examples of *local similarity measures* [68], where a local similarity measure is one that not only provides a single numeric value describing the extent of the similarity between two molecules, but that also provides an *alignment* of one molecule with another, *i.e.*, a mapping of features in the target structure to features in a database structure sufficient to superimpose one upon the other [69].

The MCS is the largest set of features (atoms, bonds or inter-atomic distances) from a target structure that can be superimposed exactly (or within user-defined tolerances) onto the database structure. Once the MCS has been found, the number of matching atoms can be substituted into a Tanimoto-like coefficient to obtain the global similarity between the target structure and each of the database structures, with the local similarity here being the alignment that gave rise to that global similarity value.

The maximum overlay of one molecule onto another is an appealing, and intuitively acceptable, measure of structural similarity (where "maximum" is defined in terms of the number of matching atoms and/or bonds in the case of 2D molecules and of atoms and/or inter-atomic distances in the case of 3D molecules). However, it does require the use of an MCS algorithm to identify the matching features in the two molecules that are to be compared. MCS detection belongs to the class of NP-complete computational problems, for which no efficient, deterministic algorithms exist and much of the Sheffield work has focused on the identification of algorithms that are sufficiently rapid to allow their use on a routine basis [70-72]; these studies have identified the Bron-Kerbosch algorithm [73], in particular, as being the most generally applicable to 3D MCS detection in chemical and biological structures. However, the very different natures of 2D and 3D chemical graphs have meant that it is only very recently that we have been able to identify an algorithm that is sufficiently fast to enable MCS-based similarity searching in 2D databases (as discussed in the final section of the paper).

While common patterns of atoms are clearly of importance in relating pairs of structures, it is generally believed that it is the electrostatic, steric and hydrophobic characteristics of molecules that are of most importance in determining their biological activities; in particular, 3D QSAR (for Quantitative Structure-Activity Relationship) is now a well-established approach to predicting the activities of molecules on the basis of the electrostatic, steric and hydrophobic fields surrounding a 3D molecule [74]. Field-based similarity measures quantify the degree of resemblance between a pair of molecules by a similarity coefficient based on the overlap of molecular 3D properties, such as the molecular electrostatic potential. A molecule is positioned at the centre of a 3D grid, the

potential calculated at each point in the grid, and then the similarity between two molecules is estimated by comparing the potentials at each grid point and summing over the entire grid (see, e.g., [75, 76]). Efficiencies in operation are obtained using the Gaussian approximation procedure of Good *et al*. [77] but there is still a need to search for the best alignment of the two molecules that are being compared, so as to ensure that analogous grid points are matched. This has meant that while field-based similarity measures are becoming widely used in QSAR (see, e.g., [78, 79]) they have traditionally been far too slow for similarity searching, where the target structure needs to be aligned with each of the database structures in turn. The need for an appropriate alignment procedure spurred a series of studies of methods for field-based similarity searching, and resulted in the development of a genetic algorithm (GA) to identify a set of translations and rotations of the target structure, relative to the database structure, required for the alignment that maximises the field-based similarity between the two molecules. The algorithm is sufficiently rapid in execution to permit the field-based searching of databases of non-trivial size [80] and has been shown to result in sets of bioactive nearest neighbours that are different in nature from those found by fragment-based 2D similarity searching [81].

The fact that we have investigated several different types of similarity measure highlights the fact that there is no single "best" way of carrying out similarity searching; instead, there is now much interest in the use of *data fusion*, where one combines the rankings resulting from several different scoring functions to give a single, combined ranking as the output of a similarity search. Data fusion (which is also referred to as *consensus scoring* when used in the context of ligand-protein docking) is now well established in textual information retrieval, where one might, for example, use a range of index term weighting schemes and then combine the rankings resulting from each of them [82]. We [83], and others [84, 85], have used the same idea to combine different measures of structural similarity for chemical similarity searching, and have found that this often results in a level of search effectiveness (however this is measured [86]) that is better than that resulting from any single similarity measure (see, e.g., [81, 87]).

5. Compound selection and library design

The widespread adoption by the pharmaceutical industry of combinatorial chemistry and high-throughput screening (HTS) from the mid-Nineties onwards has resulted in a massive increase in the numbers of compounds that can be synthesised and tested for biological activity. However, despite the increase in throughput it was quickly realised that there are many more compounds that could potentially be made than can be handled in practice (chemistry space has been estimated to contain as many as 10^{40} compounds [88]). Although, compound selection techniques had been used for many years for the selection of sets of compounds from corporate collections for biological screening, the advances in the automation of the experimental techniques led to a renewed interest in computational methods for selecting subsets of compounds. In particular, recent efforts have focused on novel methods for designing *combinatorial libraries*, where a combinatorial library is a set of compounds that can be synthesised in parallel using robotics techniques that are far faster than conventional, one-compound-at-a-time synthetic procedures. It is, however, important that an appropriate set of compounds is synthesised, and *molecular diversity analysis* is the name given to attempts to identify sets of compounds that are as *diverse* (or heterogeneous, dissimilar, widely-spaced etc.) as possible.

The rationale for diversity analysis lies in the Similar Property Principle, which has been discussed in the previous section. Given a definition of chemistry space that is relevant to biological activity, if structures that are close in the space are likely to exhibit similar bioactivity then they are likely to be redundant in terms of a biological screening experiment, as they will be unable to provide any additional information about the relationship between chemical structure and biological activity. Thus, a library that maximises coverage of structural space should also provide maximum coverage of biological activity with minimum redundancy. The main techniques that have been developed to select diverse subsets of compounds include clustering, dissimilarity-based compound selection (DBCS), partitioning or cell-based approaches, and optimisation-

based methods [89]. These methods are routinely used to select compounds for biological testing and also to select subsets of reagents for the synthesis of combinatorial libraries.

The clustering of chemical databases has been introduced in Section 4. Once a dataset has been clustered, a representative subset can be chosen by taking one compound from each cluster. Cluster-based compound selection is thus a two-stage process: in DBCS, conversely, a subset is selected directly. The basic DBCS algorithm was first described by Bawden [90] and Lajiness [91], and involves selecting the first compound at random and then iteratively choosing the next compound as the one that is most dissimilar to those that have already been selected. The Bawden-Lajiness algorithm is 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, which makes it impractical for use with large chemical databases. The selection step in each iteration involves calculating the dissimilarity of every compound remaining in the database to the compounds already selected in the subset, and the dissimilarity can be measured in different ways. For example, in the MaxSum method, it is measured as the sum of the average pairwise dissimilarities to all compounds in the subset, and in the MaxMin method it is measured as the dissimilarity of the most similar compound in the subset to the database compound [92].

The clustering of document databases was studied intensively in Sheffield in the early Eighties [8] and we were hence aware of the elegant work of Voorhees [93], who had developed an efficient way of calculating the inter-cluster similarities in the group-average method hierarchical clustering method. In this method, the inter-cluster similarity is the average of all of the pair-wise object similarities, where one document is in one of the two selected clusters and the other document is in the other cluster. Voorhees demonstrated that the similarity could be obtained from a procedure that involved just a single similarity calculation using the weighted centroids of the two clusters. We realised that this equivalence can be applied more generally to any situation where sums of similarities, rather than individual similarities, are required, providing that: the cosine coefficient is used to measure the similarity between pairs of objects; and

that the objects that are being compared are characterised by vector representations (such as fragment bit-strings). It hence proved possible to adapt Voorhees' algorithm to give a fast, $O(nN)$, implementation of the selection step of the Bawden-Lajiness algorithm using the MaxSum method [94], this implementation providing one of the first tools for large-scale DBCS.

Although we, and others, later showed that other selection algorithms are superior, [92, 95, 96] the MaxSum algorithm provided the starting point for our work on comparing reagent-based and product-based approaches to combinatorial library design. Reagent-based selection involves choosing subsets of reagents without consideration of the product molecules that will result, whereas product-based selection involves enumerating the virtual product library from all available reagents and then choosing a combinatorial subset directly from product space. Reagent-based selection is much less computationally demanding than product-based selection, but there is no guarantee that optimised subsets of reagents will lead to optimised products; in particular, we felt that reagent-based selection might result in combinatorial libraries that were less than ideally diverse. We developed a GA-based approach to product-based selection that makes use of the efficient MaxSum method for quantifying molecular diversity, and were hence able to show that product-based approaches do in fact result in more diverse libraries than reagent-based methods [97, 98], a result that was subsequently confirmed in other studies [99].

Recent work in compound selection and combinatorial library design has focused on the design of libraries optimised on a number of properties simultaneously. This is due to the poor performance of early HTS experiments where libraries either failed to deliver the improved hit rates (in terms of numbers of bioactive molecules) that were expected or resulted in hits with characteristics that made them undesirable as potential drugs. Consequently, the focus in library design has shifted towards designing libraries that are optimised on multiple properties simultaneously, for example, diversity and drug-like physicochemical properties. Most approaches to multi-objective library design are based on traditional optimisation methods such as GAs or simulated annealing that handle

multiple objectives via a weighted-sum fitness function, which effectively reduces the multiple objectives to a single value that must be maximised or minimised [100-103]. At Sheffield, we adapted our product-based approach to library design to handle multiple objectives using the weighted-sum approach. Typical objectives handled in this program, called SELECT, include diversity (or similarity to a known bioactive target structure), the cost of synthesising the library, and drug-like physicochemical properties.

Although effective in operation [100], the weighted-sum approach to optimising multiple objectives has several limitations; for example, it is usually difficult to assign appropriate weights, especially for different types of objectives such as diversity and cost, and the end result is a single compromise solution with the exact solution being determined by the relative weights. We have tackled this problem in a collaboration with colleagues in Sheffield's Department of Automatic Control and Systems Engineering on the use of a multiple objective genetic algorithm (MOGA) for library design. This provides a more effective way of combining the (often conflicting) characteristics that help to make a molecule a potential drug, and has resulted in a program, called MoSELECT, that has been designed to overcome the limitations of conventional library-design programs [104, 105]. MoSELECT is based on a MOGA in which multiple objectives are handled independently without the need to assign weights. The MOGA searches for multiple solutions in parallel and yields a family of solutions where each solution is equally valid and represents a different compromise solution to the set of constraints that are being optimised. MoSELECT thus allows the relationships between different objectives to be explored, with competing objectives being easily identified and with the user being able to make an informed choice on which solution(s) to explore.

6. Citation analysis of Sheffield chemoinformatics research

Citation analysis is widely used for assessing the influence of research groups or individuals, by consideration of the numbers of citations to their publications [106-109]. There have been, and there continue to be, many criticisms of its use for this purpose, but it does provide a relatively simple way of obtaining quantitative data on the extent to

which published work has been taken up by subsequent researchers in the field of interest. Previous studies have looked at the publications of, and citations to, members of the academic staff [110] and research students [111] in the Department. Here, we report a citation analysis of the publications of the five members of the full-time academic staff who have carried out chemoinformatics research here: George Adamson, Michael Lynch and three of the authors of this paper (VJG, JDH and PW). Complete bibliographies were available for the five individuals; after the removal of duplicate publications (i.e., those involving more than one of the authors), a total of 321 publications was identified and searched for across all three of the Web of Science databases (SCI-Expanded, SSCI and A&HCI) for the entire period for which citation data are available (1980 to date). The resulting citations were downloaded into a bibliographic management program (EndNote version 6) with subsequent analyses being carried out with Microsoft Excel.

A total of 4845 unique citations was identified to 321 publications by the five members of the Group. The figure of 4845 includes many self-citations, i.e., citations by an author to another publication by that individual: removal of these gave a total of 3725 residual citations. We wish to make three points about these figures. First, it should be noted that the figure of 3725 includes a total of 704 co-citations within the Group; thus while the residual citations do indeed quantify the extent to which the Sheffield work has influenced subsequent research, it could be argued that it over-estimates the degree of external recognition. Second, the 321 publications include non-chemoinformatics articles from the bibliographies that were searched for in the Web of Science; however, these have been included because of the close relationship that exists between chemical and textual information retrieval. This is especially true in the case of the Sheffield research [7], as exemplified by the link between document clustering and molecular diversity analysis noted in Section 4. Third, the figure of 321 publications includes 35 publications that did not attract a single citation. However, this is probably an artefact as they had all been published prior to 1980 and thus may well have had citations in the literature pre-dating the start-of-coverage of the Web of Science databases; there are also, of course, almost certainly other pre-1980 citations that are not included in the total of 4845 for the same reason. It is thus possible to conclude that each of the 321 papers received, on average, at

least 15.1 citations (or at least 11.6 residual citations). As would be expected, there is a highly skewed distribution of numbers of citations: four of the publications received at least 100 residual citations, as listed in listed in Table 1, and there were another four that received at least 75 residual citations. Of the four publications in Table 1, the first and third have already been discussed and cited in Sections 4 and 5, respectively. The second discusses an algorithm for ligand-protein docking that has since been implemented in many of the major pharmaceutical and agrochemical companies, and the fourth was the first detailed review for many years on chemical ring perception, an important graph-theoretic building block for many applications in chemoinformatics.

The journals in which the citations appeared were noted: in all, no less than 411 different journals had cited at least one of the five Sheffield authors at least once. Nine of these provided at least 100 citations but the figures here are dominated by the *Journal of Chemical Information and Computer Sciences*. This is published by the American Chemical Society and is the core journal for the field of chemoinformatics: it provided no less than 1346 citations, over one-quarter of all of the unique citations to the Group's work. Adopting Cronin and Davenport's definition of a core journal for a group of authors as being those journals that cite all members of the group at least once [112] then it is possible to identify the following eight journals as being the core for the Sheffield chemoinformatics group (they are arranged in decreasing number of citations provided): *Journal of Chemical Information and Computer Sciences*, *Journal of Computer-Aided Molecular Design*, *Journal of Molecular Graphics and Modelling*, *Annual Review of Information Science and Technology*, *Journal of Documentation*, *Drug Discovery Today*, *Combinatorial Chemistry and High Throughput Screening* and *Analytica Chimica Acta*. These are hardly surprising, covering as they do most of the major journals where chemoinformatics articles appear; the two information science journals appear in this core as they both contained review articles with many citations to work at Sheffield. What is, perhaps, more surprising is the sheer range of titles included in the 411 journals: considering just those 166 that provided a single citation, these include not only the expected chemical, chemoinformatics, biological, bioinformatics, computing and information science journals but examples from across the physical, life and medical

sciences, e.g., *Acta Alimentaria*, *Astronomy and Astrophysics*, *British Dental Journal*, *Clinical Radiology*, *Earth Surface Processes and Landforms*, *Geological Magazine*, *Journal of Immunology*, *Journal of Nutrition*, *Journal of Receptor and Signal Transduction Research*, *Nephrology Dialysis Transplantation*, *Powder Diffraction*, and *Water Environment Research.*

The citations were sub-divided on the basis of the type of organisation that published a citing paper, using the first named corporate source given in the address section of the ISI records. The largest resulting sub-division was, of course, the Department itself, accounting for all of the self-citations. The remaining unique citations were sub-divided into the following three classes: Commercial; Academic; and Other. Commercial sources were identified by the appearance of terms such as "Co.", "Ltd." and "Inc." within the address details. Academic sources were identified by the appearance of terms such as "University", "Institute" or "Academy"; individual departments within the same academic source were regarded as distinct. The "Other" group contained a range of independent and governmental research laboratories, together with the sources from 32 papers that did not provide sufficient information to identify the organisations that undertook the research. A total of 910 different organisations, excluding the Department, cited the set of Sheffield publications, the distribution being as shown in Table 2, which also lists *k*, the mean citation count per corporate source. In considering these figures it should be remembered that it is not uncommon, in the case of academic and company collaborations, for the academic partner to appear first in the list; the "Academic" figures may thus have been boosted, and the "Commercial" figures reduced, for this reason.

Considering first the "Commercial" sources, the most frequently citing organisation was Barnard Chemical Information Limited, a specialist chemical software company that employs several past research staff and research students from the Group; other frequent citers include Chemical Abstracts, Proteus Molecular Design Limited (another chemical software company), and many of the major pharmaceutical companies (including such multi-nationals as Bristol Myers Squibb, GlaxoWellcome, Novartis and Pfizer). Although there are many less commercial than non-commercial organisations, the *k*

values in the right-hand column of Table 2 demonstrate that the former make very extensive use of the Sheffield research. The most frequently citing "Academic" organisation was the Cambridge Crystallographic Data Centre at the University of Cambridge, with which we have had several successful collaborations (see, e.g., [65, 113]); other frequent citers include the chemoinformatics groups at the Universities of California at San Francisco, Leeds and Paris. The most frequently citing "Other" organisation was the National Cancer Institute (part of the National Institutes of Health in the USA); other frequent citers included the European Molecular Biology Laboratory, the Biomolecular Modelling Laboratory of the Imperial Cancer Research Fund, and the Institute for Algorithms and Scientific Computing of the German National Research Centre for Information Technology. We also noted the countries of origin of the first-named citing organisation. As might be expected, the bulk of the citations came from the UK and from the USA, these providing 867 residual citations and 1550 residual citations, respectively, with the remaining 1308 residual citations coming from a total of 52 different countries. This again demonstrates the breadth of influence of the Group's research.

7. Conclusions

In this paper, we have provided a brief overview of the chemoinformatics research that has been carried out in the Department of Information Studies at the University of Sheffield, focusing principally on research since 1985. The work summarised here has involved the development and testing of algorithms for a range of applications in chemoinformatics, including 3D substructure searching, Markush patent searching, 2D and 3D similarity searching, and molecular diversity analysis. However, it must be emphasised that there have been several other areas that have attracted serious attention during the period under review: for example, we have not mentioned work on pharmacophore mapping, ligand docking and 3D QSAR, for all of which there is now widely used commercial software that draws on the Sheffield work (see, e.g., [113-115]).

Of the work discussed here, that on 3D substructure searching and on Markush searching is now complete, but extensive studies continue in other areas. For example, in the similarity area, we have recently described a new algorithm [116] that is sufficiently fast to allow MCS-based searching of large 2D databases, something that has not previously been possible; our initial results suggest that the rankings resulting from this algorithm are very different from conventional fragment-based similarity searches, with the best results being obtained by fusing the two types of search output [117]. In the bioinformatics area, we have started to apply our graph-theoretic methods for searching 3D proteins [10] to the representation and searching of RNA structures; this work provides the first systematic approach to the retrieval of user-defined patterns of bases and has already been shown to identify previously unknown occurrences of such patterns [118]. Finally, in the compound selection area, we are looking at ways in which we can apply the reduced graph concept that was first developed for the representation and searching of generic chemical structures [44]. Reduced graphs are being used to encode the potential pharmacophore points in molecules, but using the molecule's topology rather than its geometry (as in the work described in Section 2 on 3D substructure searching). Our initial results suggest that, given an appropriate level of graph reduction, this may provide a simple and effective way of probing the biological activities of sets of compounds [119].

Acknowledgements. We thank all the members, past and present, of the chemoinformatics research group for their contributions; in particular, we thank Michael Lynch, who initiated work on chemoinformatics in Sheffield and who acted as the PhD supervisor for three of us (VJG, JDH and PW). Funding for the studies reported here has been provided by AstraZeneca, the Biotechnology and Biological Sciences Research Council, the British Library Research and Development Department, the Cambridge Crystallographic Data Centre, Chemical Abstracts Service, Derwent Publications, Eli Lilly, Elsevier, the Engineering and Physical Sciences Research Council, GlaxoSmithKline, GlaxoWellcome, ICI Agrochemicals, ICI Pharmaceuticals, International Documentation in Chemistry, the James Black Foundation, the Medical Research Council, Novartis, Oxford Asymmetry International, Pfizer, the Science and Engineering Research Council, Syngenta, Tripos, Unilever, Warner-Lambert Parke-Davis and Zeneca Agrochemicals. Hardware, software, database and laboratory support have been provided by Barnard Chemical Information, Cambridge Crystallographic Data Centre, Current Drugs Limited, Daylight Chemical Information Systems, MDL Information Systems, the Royal Society, Synopsys Scientific Systems, Tripos, and the Wolfson Foundation.

References

1. M. Hann and R. Green, Chemoinformatics – a new name for an old problem? *Current Opinion in Chemical Biology*, 3 (1999) 379-383.
2. L.C. Ray and R.A. Kirsch, Finding chemical records by digital computers. *Science*, 126 (1957) 814-819.
3. G.M. Dyson and M.F. Lynch, *Chemical-Biological Activities* – a computer-produced express digest. *Journal of Chemical Documentation*, 3 (1963) 81-85.
4. W.E. Cossum, M.L. Krakiwsky and M.F. Lynch, Advances in automatic chemical substructure searching techniques. *Journal of Chemical Documentation*, 5 (1965) 33-35.
5. M.F. Lynch, Subject indexes and automatic document retrieval – the structure of index entries in Chemical Abstracts subject indexes. *Journal of Documentation*, 22 (1966) 167-185.
6. J.E. Armitage, J.E. Crowe, P.N. Evans, M.F. Lynch and J.A. McGuirk, Documentation of chemical reactions by computer analysis of structural changes. *Journal of Chemical Documentation*, 7 (1967) 209-215.
7. P. Willett, Textual and chemical information retrieval: different applications but similar algorithms. *Information Research*, 5(2) (2000) at URL http://InformationR.net/ir/5-2/infres52.html
8. M.F. Lynch and P. Willett, Information retrieval research in the Department of Information Studies, University of Sheffield: 1965-1985. *Journal of Information Science*, 13 (1987) 221-234.
9. P. J. Artymiuk, R.V. Spriggs and P. Willett, Graph theoretic methods for the analysis of structural relationships in biological macromolecules. Submitted for publication.
10. J.M. Barnard, Substructure searching methods: old and new. *Journal of Chemical Information and Computer Sciences*, 33 (1993) 532-538.
11. P. Gund, Three-dimensional pharmacophoric pattern searching. *Progress in Molecular and Subcellular Biology*, 5 (1977) 117-143.
12. K.Y. Zee-Cheng and C.C. Cheng, Common receptor-complement feature among some anti-leukemic compounds. *Journal of Pharmaceutical Science*, 59 (1970) 1630-1634.
13. S.E. Jakes and P. Willett, Pharmacophoric pattern matching in files of 3-D chemical structures: selection of inter-atomic distance screens. *Journal of Molecular Graphics*, 4 (1986) 12-20.
14. J.K. Cringean, C.A. Pepperrell, A.R. Poirrette and P. Willett, Selection of screens for three-dimensional substructure searching. *Tetrahedron Computer Methodology*, 3 (1990) 37-46.
15. R.P. Sheridan, R. Nilakantan, A. Rusinko, N. Bauman, K.S. Haraki and R. Venkataraghavan, 3DSEARCH: a system for three-dimensional substructure searching. *Journal of Chemical Information and Computer Sciences*, 29 (1989) 255-260.
16. W. Fisanick, K.P. Cross and A. Rusinko, Similarity searching on CAS Registry substances. 1. Global molecular property and generic atom triangle geometric searching. *Journal of Chemical Information and Computer Sciences*, 32 (1992) 664-674.
17. S.E. Jakes, N.J. Watts, P. Willett, D. Bawden and J.D. Fisher, Pharmacophoric pattern matching in files of 3-D chemical structures: evaluation of search performance. *Journal of Molecular Graphics*, 5 (1987) 41-48.
18. A.C. Good and J.S. Mason, Three-dimensional structure database searches. *Reviews in Computational Chemistry*, 7 (1996) 67-117.
19. A.T. Brint and P. Willett, Pharmacophoric pattern matching in files of 3-D chemical structures: comparison of geometric searching algorithms. *Journal of Molecular Graphics*, 5 (1987) 49-56.
20. J.R. Ullmann, An algorithm for subgraph isomorphism *Journal of the ACM*, 16 (1976) 31-42.
21. W.A. Warr and P. Willett, The principles and practice of 3D database searching. In: Y.C. Martin and P. Willett (eds.), *Designing Bioactive Molecules: Three-Dimensional Techniques and Applications* (American Chemical Society, Washington DC, 1997).
22. G.M. Crippen and T.F. Havel, *Distance Geometry and Molecular Conformation* (Research Studies Press, Letchworth, 1988).
23. D.E. Clark, P. Willett and P.W. Kenny, Pharmacophoric pattern matching in files of three-dimensional chemical structures: use of smoothed bounded-distance matrices for the representation and searching of conformationally-flexible molecules. *Journal of Molecular Graphics*, 10 (1992) 194-204.
24. D.E. Clark, G. Jones, P. Willett, P.W. Kenny and R.C. Glen, Pharmacophoric pattern matching in files of three-dimensional chemical structures: comparison of conformational-searching algorithms for flexible searching. *Journal of Chemical Information and Computer Sciences*, 34 (1994) 197-206.

25. T. Hurst, Flexible 3D searching: the directed tweak technique. *Journal of Chemical Information and Computer Sciences*, 34 (1994) 190-196.
26. M.G. Bures, Integration of molecular modelling and database searching. In: Y.C. Martin and P. Willett (eds.), *Designing Bioactive Molecules: Three-Dimensional Techniques and Applications* (American Chemical Society, Washington DC, 1997).
27. T.E. Moock, D.R. Henry, A.G. Ozkaback and M. Alamgir, Conformational searching in ISIS/3D databases. *Journal of Chemical Information and Computer Sciences*, 34 (1994) 184-189.
28. The UNITY software is available from Tripos Inc. at http://www.tripos.com
29. J.F.B Rowland, *Information Transfer and Use in Chemistry*, British Library Research and Development Department Report No. 5385 (British Library Research and Development Department, London, 1978).
30. W. Dethlefsen, M.F. Lynch, V.J. Gillet, G.M. Downs, J.D. Holliday and J.M. Barnard, Computer storage and retrieval of generic chemical structures in patents, Part 11. Theoretical aspects of the use of structure languages in a retrieval system. *Journal of Chemical Information and Computer Sciences*, 31 (1991) 233-253.
31. W. Dethlefsen, M.F. Lynch, V.J. Gillet, G.M. Downs, J.D. Holliday and J.M. Barnard, Computer storage and retrieval of generic chemical structures in patents, Part 12. Principles of search operations involving parameter lists: matching-relations, user-defined match levels, and transition from the reduced graph to the refined search. *Journal of Chemical Information and Computer Sciences*, 31 (1991) 253-260.
32. W. Fisanick, The Chemical Abstracts Service generic chemical (Markush) storage and retrieval capability, Part 1. Basic concepts. *Journal of Chemical Information and Computer Sciences*, 30 (1990) 145-155.
33. T. Ebe, K.A. Sanderson and P.S. Wilson, The Chemical Abstracts Service generic chemical (Markush) storage and retrieval capability, Part 2. The MARPAT file. *Journal of Chemical Information and Computer Sciences*, 31 (1991) 31-36.
34. K. Shenton, P. Norton, and E.A. Fearns, Generic searching of patent Information. In: W.A. Warr (ed.), *Chemical Structures: The International Language of Chemistry* (Springer, Berlin, 1988).
35. M.F. Lynch and J.D. Holliday, The Sheffield Generic Structures Project – a retrospective review. *Journal of Chemical Information and Computer Sciences*, 36 (1996) 930-936.
36. J.M. Barnard, M.F. Lynch and S.M. Welford, Computer storage and retrieval of generic chemical structures in patents, Part 2. GENSAL: a formal language for the description of generic chemical structures. *Journal of Chemical Information and Computer Sciences*, 21 (1981) 151-161.
37. J.M. Barnard, M.F. Lynch and S.M. Welford, S.M. Computer storage and retrieval of generic chemical structures in patents, Part 6. An interpreter program for the generic structure language GENSAL. *Journal of Chemical Information and Computer Sciences*, 24 (1984) 66-70.
38. J.M. Barnard, M.F. Lynch and S.M. Welford, Computer storage and retrieval of generic chemical structures in patents, Part 4. An extended connection table representation (ECTR) for generic structures. *Journal of Chemical Information and Computer Sciences*, 22 (1982) 160-164.
39. P.G. Dittmar, N.A. Farmer, W. Fisanick, R.C. Haines, J.A. Miller and B. Koch, The CAS ONLINE search system. I. General system design and selection, generation and use of search screens. *Journal of Chemical Information and Computer Sciences*, 33 (1983) 93-102.
40. G.M. Downs, V.J. Gillet, J.D. Holliday and M.F. Lynch, Computer storage and retrieval of generic chemical structures in patents, Part 9. An algorithm to find the extended set of smallest rings in structurally explicit generics. *Journal of Chemical Information and Computer Sciences*, 29 (1989) 207-214.
41. J.D. Holliday, V.J. Gillet, G.M. Downs, M.F. Lynch and W. Dethlefsen, Computer storage and retrieval of generic chemical structures in patents, Part 14. Algorithmic generation of fragment descriptors for generic structures. *Journal of Chemical Information and Computer Sciences*, 32 (1992) 453-462.
42. G.M. Downs, V.J. Gillet, J.D. Holliday and M.F. Lynch, Computer storage and retrieval of generic chemical structures in patents, Part 10. Assignment and logical bubble-up of ring screens for structurally explicit generics. *Journal of Chemical Information and Computer Sciences*, 29 (1989) 215-224.
43. J.D. Holliday, G.M. Downs, V.J. Gillet and M.F. Lynch, Computer storage and retrieval of generic chemical structures in patents, Part 15. Generation of topological fragment descriptors from

nontopological representations of generic structure components. *Journal of Chemical Information and Computer Sciences*, 33 (1993) 369-377.

44. V.J. Gillet, G.M. Downs, A. Ling, M.F. Lynch, P. Venkataram, J.V. Wood and W. Dethlefsen, Computer storage and retrieval of generic chemical structures in patents, Part 8. Reduced chemical graphs and their applications in generic chemical structure retrieval. *Journal of Chemical Information and Computer Sciences*, 27 (1987) 126-137.
45. J.D. Holliday and M.F. Lynch, Computer storage and retrieval of generic chemical structures in patents, Part 16. The refined search: an algorithm for matching components of generic chemical structures at the atom-bond level. *Journal of Chemical Information and Computer Sciences*, 35 (1995) 1-7.
46. K. Sparck Jones and P. Willett (eds.), *Readings in Information Retrieval*. (Morgan Kaufmann, San Francisco CA, 1997).
47. R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. (Addison Wesley, Harlow, 1999).
48. M.A. Johnson and G.M. Maggiora (eds.), *Concepts and Applications of Molecular Similarity* (John Wiley, New York, 1990).
49. C.J. van Rijsbergen, *Information Retrieval*. (Butterworth, London, 1979).
50. G.W. Adamson and J. A. Bush, A method for the automatic classification of chemical structures. *Information Storage and Retrieval*, 9 (1973) 561-568.
51. R.E. Carhart, D.H. Smith and R. Venkataraghavan, Atom pairs as molecular features in structure-activity studies: definition and application. *Journal of Chemical Information and Computer Sciences*, 25 (1985) 64-73.
52. P. Willett, V. Winterman and D. Bawden, Implementation of nearest neighbour searching in an online chemical structure search system. *Journal of Chemical Information and Computer Sciences*, 26 (1986) 36-41.
53. P.M. Dean (ed.), *Molecular Similarity in Drug Design* (Chapman and Hall, Glasgow, 1994).
54. P. Willett, J.M. Barnard and G.M. Downs, Chemical similarity searching. *Journal of Chemical Information and Computer Sciences*, 38 (1998) 983-996.
55. P.H.A. Sneath, Relations between chemical structure and biological activity in peptides. *Journal of Theoretical Biology*, 12 (1966) 157-195.
56. P.J. Harrison, A method of cluster analysis and some applications. *Applied Statistics*, 17 (1968) 226-236.
57. P. Willett, *Similarity and Clustering in Chemical Information Systems* (Research Studies Press, Letchworth, 1987).
58. R.A. Jarvis and E.A. Patrick, Clustering using a similarity measure based on shared nearest neighbours. *IEEE Transactions on Computers*, C-22 (1973) 1025-1034.
59. P. Willett, V. Winterman and D. Bawden, Implementation of non-hierarchic cluster analysis methods in chemical information systems: selection of compounds for biological testing and clustering of substructure search output. *Journal of Chemical Information and Computer Sciences*, 26 (1986) 109-118.
60. N.E. Shemetulskis, J.B. Dunbar, B.W. Dunbar, D.W. Moreland and C. Humblet, Enhancing the diversity of a corporate database using chemical database clustering and analysis. *Journal of Computer-Aided Molecular Design*, 9 (1995) 407-416.
61. G.M. Downs, P. Willett and W. Fisanick, Similarity searching and clustering of chemical-structure databases using molecular property data. *Journal of Chemical Information and Computer Sciences*, 34 (1994) 1094-1102.
62. R.D. Brown and Y.C. Martin, 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 (1996) 572-584.
63. J.H. Ward, Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 58 (1963) 236-244.
64. C.A. Pepperrell and P. Willett, Techniques for the calculation of three-dimensional structural similarity using inter-atomic distances. *Journal of Computer-Aided Molecular Design*, 5 (1991) 455-474.

65. P.A. Bath, A.R. Poirrette, P. Willett and F.H. Allen, Similarity searching in files of three-dimensional chemical structures: comparison of fragment-based measures of shape similarity. *Journal of Chemical Information and Computer Sciences*, 34 (1994) 141-147.
66. A.C. Good, T.J.A. Ewing, D.A. Gschwend and I.D. Kuntz, New molecular shape descriptors – application in database screening. *Journal of Computer-Aided Molecular Design*, 9 (1995) 1-12.
67. R.P. Sheridan, M.D. Miller, D.J. Underwood and S.K. Kearsley, Chemical similarity using geometric atom pair descriptors. *Journal of Chemical Information and Computer Sciences*, 36 (1996) 128-136.
68. G.M. Downs and P. Willett, Similarity searching in databases of chemical structures. *Reviews in Computational Chemistry*, 7 (1995) 1-66.
69. C. Lemmen and T. Lengauer, Computational methods for structural alignment of molecules. *Journal of Computer-Aided Molecular Design*, 14 (2000) 215-232.
70. A.T. Brint and P. Willett, Algorithms for the identification of three-dimensional maximal common substructures. *Journal of Chemical Information and Computer Sciences*, 27 (1987) 152-158.
71. E.J. Gardiner, P.J. Artymiuk and P. Willett, Clique-detection algorithms for matching three-dimensional molecular structures. *Journal of Molecular Graphics and Modelling*, 15 (1998) 245-253.
72. E.J. Gardiner, J.D. Holliday, P. Willett, D.J. Wilton and P.J. Artymiuk, Selection of reagents for combinatorial synthesis using clique detection. *Quantitative Structure-Activity Relationships*, 17 (1998) 232-236.
73. C. Bron and J. Kerbosch, Algorithm 457. Finding all cliques of an undirected graph. *Communications of the ACM*, 16 (1973) 575-577.
74. H. Kubinyi, G. Folkers and Y.C. Martin (eds.) *3D QSAR in Drug Design* (Kluwer/ESCOM, Leiden, 1998).
75. R. Carbo, L. Leyda and M. Arnau, An electron density measure of the similarity between two compounds. *International Journal of Quantum Chemistry*, 17 (1980) 1185-1189.
76. M. Manaut, F. Sanz, J. Jose and M. Milesi, Automatic search for maximum similarity between molecular electrostatic potential distributions. *Journal of Computer-Aided Molecular Design*, 5 (1991) 371-380.
77. A.C. Good, E.E. Hodgkin and W.G. Richards, The utilization of Gaussian functions for the rapid evaluation of molecular similarity. *Journal of Chemical Information and Computer Sciences*, 32 (1992) 188-191.
78. A.C. Good, S.J. Peterson and W.G. Richards, QSAR's from similarity matrices. Technique validation and application in the comparison of different similarity evaluation methods. *Journal of Medicinal Chemistry*, 36 (1993) 2929-2937.
79. J. Mestres, D.C. Rohrer and G.M. Maggiora, A molecular-field-based similarity study of non-nucleoside HIV-1 reverse transcriptase inhibitors. 2. The relationship between alignment solutions obtained from conformationally rigid and flexible matching. *Journal of Computer-Aided Molecular Design*, 14 (2000) 39-51.
80. D.J. Wild and P. Willett, Similarity searching in files of three-dimensional chemical structures: alignment of molecular electrostatic potentials with a genetic algorithm. *Journal of Chemical Information and Computer Sciences*, 36 (1996) 159-167.
81. A. Schuffenhauer, V.J. Gillet and P. Willett, Similarity searching in files of 3D chemical structures: analysis of the BIOSTER database using 2D fingerprints and molecular field descriptors. *Journal of Chemical Information and Computer Sciences*, 40 (2000) 295-307.
82. N.J. Belkin, P. Kantor, E.A. Fox and J.A. Shaw, Combining the evidence of multiple query representations for information retrieval. *Information Processing and Management*, 31 (1995) 431-448.
83. C.M.R. Ginn, P. Willett and J. Bradshaw, Combination of molecular similarity measures using data fusion. *Perspectives in Drug Discovery and Design*, 20 (2000), 1-16.
84. S.K. Kearsley, S. Sallamack, E.M. Fluder, J.D. Andose, R.T. Mosley and R.P. Sheridan, Chemical similarity using physicochemical property descriptors. *Journal of Chemical Information and Computer Sciences*, 36 (1996) 118-127.
85. P.S. Charifson, J.J. Corkery, M.A. Murcko and W.P. Walters, Consensus scoring: a method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *Journal of Medicinal Chemistry*, 42 (1999) 5100-5109.

86. S.J. Edgar, J.D. Holliday and P. Willett, Effectiveness of retrieval in similarity searches of chemical databases: a review of performance measures. *Journal of Molecular Graphics and Modelling*, 18 (2000) 343-357.
87. J.D. Holliday, C.-Y. Hu and P. Willett, Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D fragment bit-strings. *Combinatorial Chemistry and High-Throughput Screening*, 5 (2002) 155-166.
88. M.J. Valler and D. Green, Diversity screening versus focussed screening in drug discovery, *Drug Discovery Today*, 5 (2000) 286-293.
89. P.M. Dean and R.A. Lewis (eds.), *Molecular Diversity in Drug Design* (Kluwer, Amsterdam, 1999).
90. D. Bawden, Molecular dissimilarity in chemical information systems. In: W.A. Warr (ed.), *Chemical Structures 2* (Springer-Verlag, Heidelberg, 1993).
91. M.S. Lajiness, Molecular similarity-based methods for selecting compounds for screening. In: D.H. Rouvray (ed.), *Computational Chemical Graph Theory* (New York: Nova Science Publishers, New York, 1990).
92. M. Snarey, N.K. Terret, P. Willett, and D.J. Wilton, Comparison of algorithms for dissimilarity-based compound celection. *Journal of Molecular Graphics and Modelling*, 15, (1998) 372-385.
93. E.M. Voorhees, Implementing agglomerative hierarchic clustering algorithms for use in document retrieval. *Information Processing and Management*, 22 (1986) 465-476.
94. J.D. Holliday, S.S. Ranade and P. Willett, A fast algorithm for selecting sets of dissimilar molecules from large chemical databases. *Quantitative Structure-Activity Relationships*, 14 (1995) 501-506.
95. D. Agrafiotis and V.S. Lobanov, An efficient implementation of distance-based diversity measures based on K-D Trees. *Journal of Chemical Information and Computer Sciences*, 39 (1999) 51-58.
96. J. Mount, J. Ruppert, W. Welch and A.N. Jain, Icepick: a flexible surface-based system for molecular diversity. *Journal of Medicinal Chemistry*, 42 (1999) 60-66.
97. V.J. Gillet, P. Willett and J. Bradshaw, The effectiveness of reactant pools for generating structurally-diverse combinatorial libraries. *Journal of Chemical Information and Computer Sciences*, 37 (1997) 731-740.
98. V.J. Gillet and O, Nicolotti, New algorithms for compound selection and library design. *Perspectives in Drug Discovery and Design*, 20 (2000) 265-287.
99. E.A. Jamois, M. Hassan and M. Waldman, Evaluation of reagent-based and product-based strategies in the design of combinatorial libraries. *Journal of Chemical Information and Computer Sciences*, 40 (2000) 63-70.
100. V.J. Gillet, P. Willett, J. Bradshaw and D.V.S Green, Selecting combinatorial libraries to optimize diversity and physical properties. *Journal of Chemical Information and Computer Sciences*, 39 (1999) 169-177.
101. R.D. Brown and Y.C. Martin, Designing combinatorial library mixtures using a genetic algorithm, *Journal of Medicinal Chemistry*, 40 (1997) 2304-2313.
102. D.N. Rassokhin, and D.K. Agrafiotis, Kolmogorov-Smirnov statistic and its application in library design, *Journal of Molecular Graphics and Modelling*, 18 (2000) 427-437.
103. J.D. Brown, M. Hassan and M. Waldman, Combinatorial library design for diversity, cost efficiency, and drug-like character. *Journal of Molecular Graphics and Modelling*, 18 (2000) 427-437.
104. V.J. Gillet, W. Khatib, P. Willett, P.J. Fleming and D.V.S. Green, Combinatorial library design using a multiobjective genetic algorithm. *Journal of Chemical Information and Computer Sciences*. 42 (2002) 375-385.
105. V.J. Gillet, P. Willett, P.J. Fleming and D.V.S. Green, Designing focused libraries using MoSELECT. *Journal of Molecular Graphics and Modelling*, 20 (2002) 491-498.
106. E. Garfield, *Citation Indexing: its theory and application in science, technology and humanities* (John Wiley, New York, 1979).
107. B. Cronin, *The Citation Process: the Role and Significance of Citations in Scientific Communication* (Taylor Graham, London, 1984).
108. M. Liu, The complexities of citation practice: a review of citation studies. *Journal of Documentation*, 49 (1993) 370-408.
109. L.M. Baird and C. Oppenheim, Do citations matter? *Journal of Information Science*, 20 (1994) 2-15.

110. S.J. Bradley, P. Willett and F.E. Wood, A publication and citation analysis of the Department of Information Studies, University of Sheffield, 1980-1990. *Journal of Information Science*, 18 (1992) 225-232.
111. M. Santos, P. Willett and F.E. Wood, Research degrees in librarianship and information science: a survey of master's and doctoral students from the Department of Information Studies, University of Sheffield. *Journal of Librarianship and Information Science*, 30 (1998) 49-56.
112. B. Cronin and L. Davenport, Profiling the professors. *Journal of Information Science*, 15 (1989) 13-20.
113. G. Jones, P. Willett, R.C. Glen, A.R. Leach and R. Taylor, Development and validation of a genetic algorithm for flexible docking. *Journal of Molecular Biology*, 267 (1997) 727-748.
114. G. Jones, P. Willett and R.C. Glen, A genetic algorithm for flexible molecular overlay and pharmacophore detection. *Journal of Computer-Aided Molecular Design*, 9 (1995) 532-549.
115. D.B. Turner, P. Willett, A.M. Ferguson and T.W. Heritage, Evaluation of a novel infra-red range vibration-based descriptor (EVA) for QSAR studies. I: General application. *Journal of Computer-Aided Molecular Design*, 11 (1997) 409-422.
116. J.W. Raymond, E.J. Gardiner and P. Willett, RASCAL: Calculation of graph similarity using maximum common edge subgraphs. *Computer Journal*, in the press.
117. J.W. Raymond and P. Willett, Effectiveness of graph-based and fingerprint-based similarity measures for virtual screening of 2D chemical structure databases. *Journal of Computer-Aided Molecular Design*, 16 (2002) 59-71.
118. A.-M. Harrison, D.R. South, P. Willett and P.J. Artymiuk, Representation and searching of patterns of bases in complex RNA structures. Submitted for publication.
119. V.J. Gillet, P. Willett and J. Bradshaw, Reduced graphs as descriptors of bioactivity. Submitted for publication.

Publication Details	Residual Citations
Willett, P. *Similarity and Clustering in Chemical Information Systems* (Research Studies Press, Letchworth, 1987)	152
G. Jones, P. Willett, R.C. Glen, A.R. Leach and R. Taylor, Development and validation of a genetic algorithm for flexible docking. *Journal of Molecular Biology*, 267 (1997) 727-748.	141
V.J. Gillet, P. Willett and J. Bradshaw, The effectiveness of reactant pools for generating structurally diverse combinatorial libraries. *Journal of Chemical Information and Computer Sciences*, 37 (1997) 731-740.	122
G.M. Downs, V.J. Gillet, J.D. Holliday and M.F. Lynch, Review of ring perception algorithms for chemical graphs. *Journal of Chemical Information and Computer Sciences*, 29 (1989) 172-187.	100

Table 1. Group publications attracting at least 100 residual citations

Organisational Type	Number Of Sources	Number Of Citations	*k*
Commercial	162	1337	8.3
Academic	617	1975	3.2
Other	131	293	2.2

Table 2. Types of citing organisation

Captions for figures

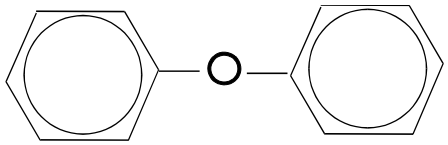
Figure 1. Example of a 2D substructure query and five of the hits resulting from a search of the NCI database.

Figure 2. Example of a 3D substructure search query.

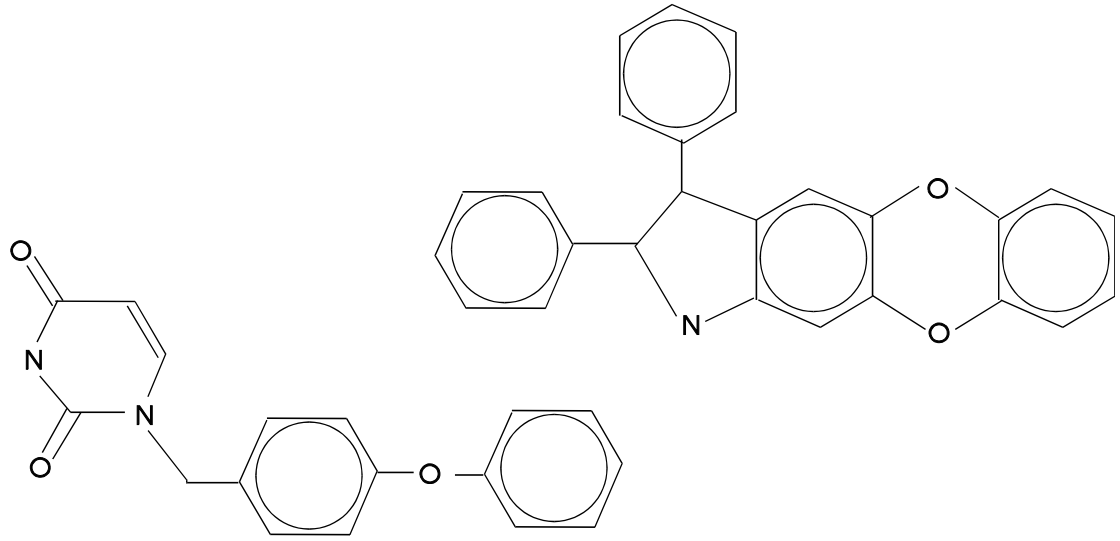
Figure 3. Five of the hits resulting from a rigid 3D search of the NCI database. The matched atoms are linked by dotted lines.

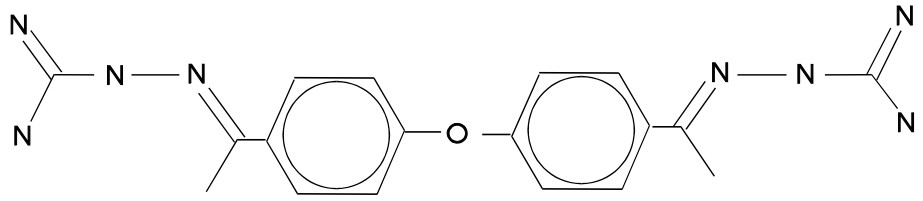
Figure 4. Example of a generic chemical structure.

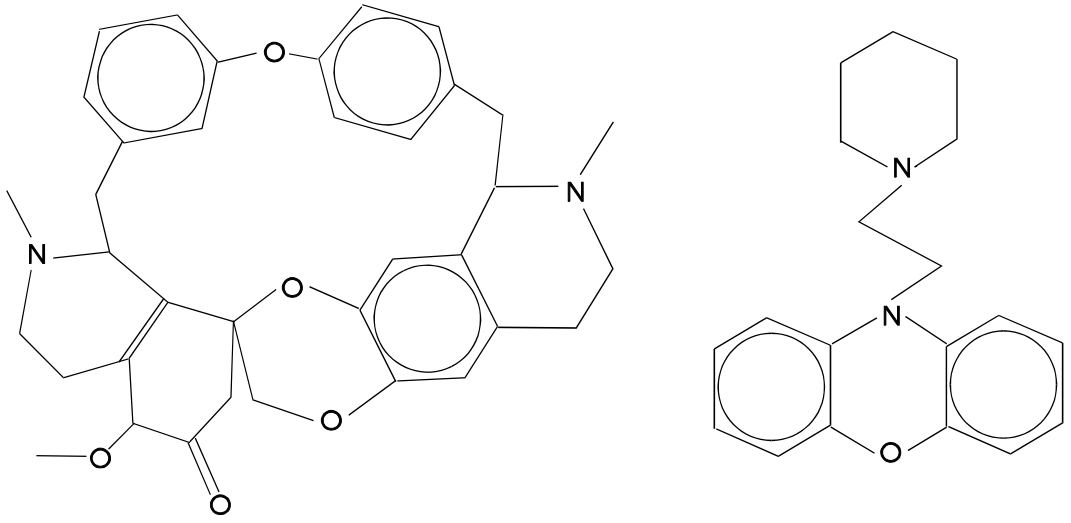
Figure 5. Example of a 2D similarity search, showing a query molecule and five of its nearest neighbours. The similarity measure is based on fragment bit-strings and the Tanimoto coefficient.

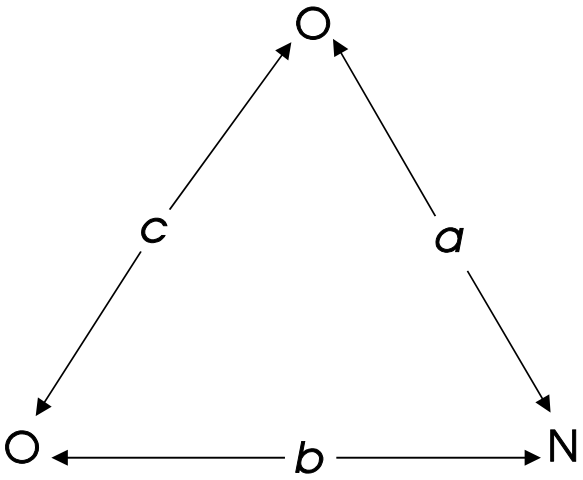


Query

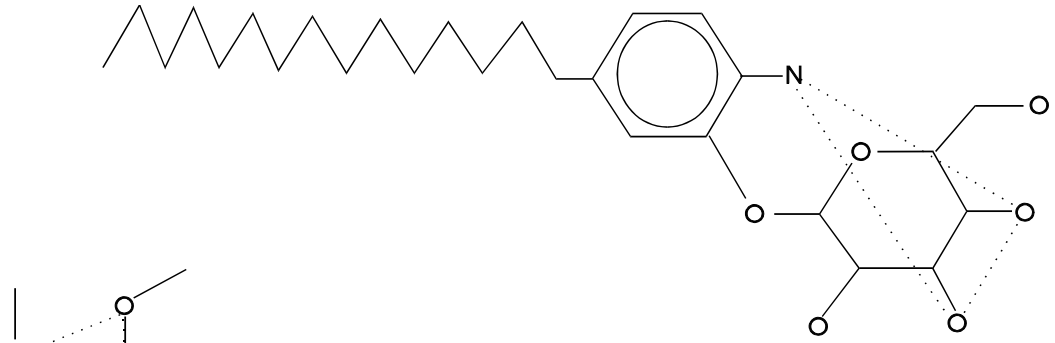


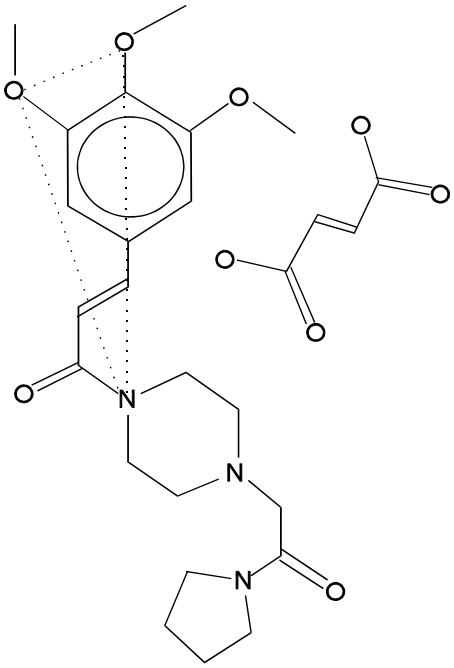


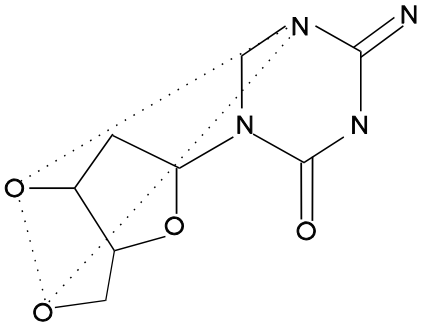


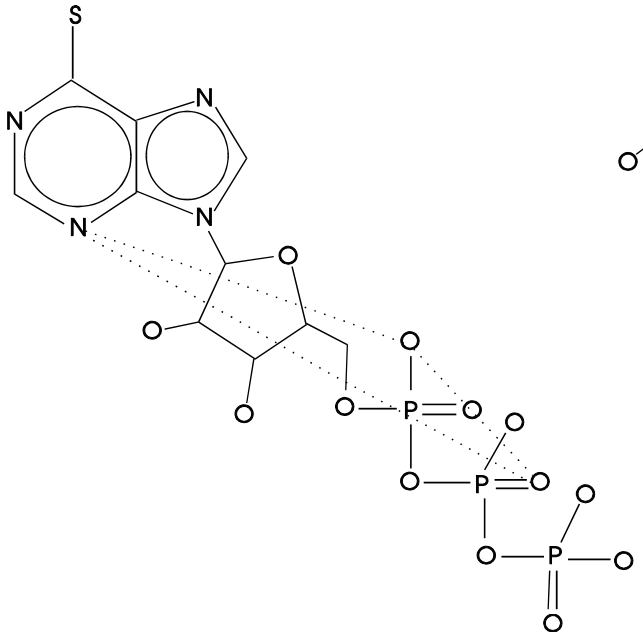


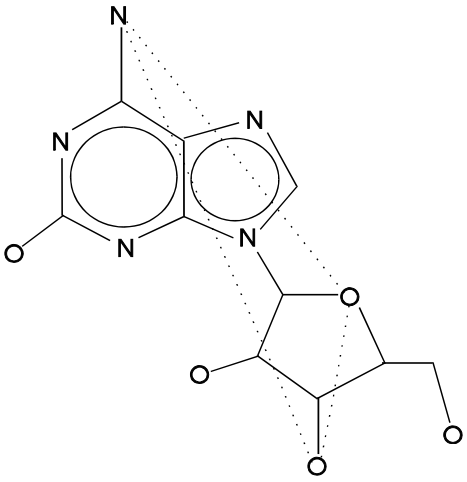
a : 8.62 ± 0.58 Å
b : 7.08 ± 0.56 Å
c : 3.35 ± 0.65 Å

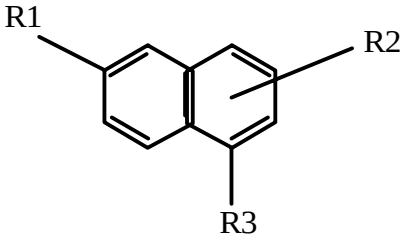




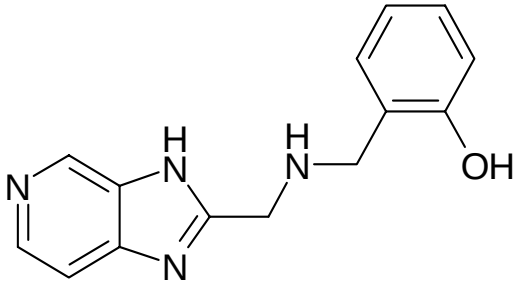


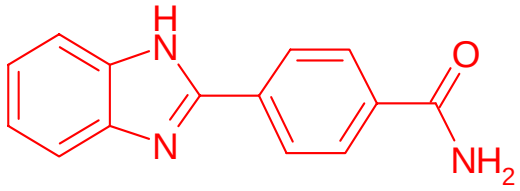






R1 is H, Cl or $(CH_2)_nCH_3$
n is 2 to 4
R2 is F or Cl
R3 is 1-3 carbon alkyl, an oxygen-containing ring or an electron withdrawing group.





Query