Dissimilarity-Based Algorithms For Selecting Structurally Diverse Sets Of Compounds

PETER WILLETT

Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield, Western Bank, Sheffield S10 2TN, UK

ABSTRACT

This paper commences with a brief introduction to modern techniques for the computational analysis of molecular diversity and the design of combinatorial libraries. It then reviews dissimilarity-based algorithms for the selection of structurally diverse sets of compounds in chemical databases. Procedures are described for selecting a diverse subset of an entire database, and for selecting diverse combinatorial libraries using both reagent-based and product-based selection.

Key words: combinatorial chemistry, dissimilarity-based compound selection, library design, molecular diversity

1. MOLECULAR DIVERSITY ANALYSIS

The pharmaceutical industry makes extensive use of highly sophisticated systems for the processing of chemical structure information (Ash *et al.*, 1991). A chemical structure diagram is represented by a graph, whose nodes and edges denote the atoms and the bonds, respectively, of a molecule; graph representations can also be used for the representation and searching of databases of 3D structures (Martin and Willett, 1998). This representation enables a range of database searching facilities to be provided by means of graph isomorphism algorithms that provide an effective, and surprisingly efficient, way of identifying molecules from a database that satisfy user-defined structural queries, *e.g.*, the retrieval of all molecules that contain a penicillin ring system or of those molecules that are most similar to a known drug. These database-searching methods are now increasingly being used to support programmes in *combinatorial chemistry* (Chaiken and Janda, 1996; DeWitt and Czarnik, 1997). This is the name given to 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 the library, *i.e.*, the degree of structural variation that is present within the set of product molecules resulting from a combinatorial synthesis. There has thus been much interest in the development of computational tools for maximising chemical diversity, especially as the techniques that have been developed are also applicable to related tasks such as the identification of structural overlap in databases and the mapping of structural space (Dean and Lewis, 1999).

The concept of diversity is normally quantified using techniques derived from those developed for *similarity searching* in chemical databases (Downs and Willett, 1995). Similarity searching involves comparing the set of structural descriptors that characterise a user-defined target structure (typically a molecule that has been shown previously to exhibit activity in a biological test) with the corresponding sets of descriptors for each of the database structures. Each such comparison results in the calculation of a measure of inter-molecular structural similarity. The similarity scores are sorted to give a ranked list in which the structures that the system judges to be most similar to the target structure, the *nearest neighbours*, are displayed first to the user. Descriptors for similarity and diversity studies are reviewed by Brown (1997): thus far, the two most important types have been *fragment substructures* and *physical properties*. In the former case, a molecule is checked for the presence of various atom- or bond-centred fragment substructures and their presence encoded in a bit-string vector, or *fingerprint*. The similarity between a pair of molecules is then calculated by a simple comparison of their associated fingerprints; alternatively, a molecule can be characterised by calculating a set of physical properties that describe its topological, electronic, steric, lipophilic or geometric features. Many coefficients are available for the calculation of inter-molecular similarities based on such descriptors (Willett *et al.*, 1998)

Considerations of cost-effectiveness dictate that as few compounds as possible should be selected for synthesis and biological testing while still ensuring coverage of the full range of structural types that are present in a dataset. There is a trivial algorithm available to identify the most

diverse *n*-compound subset of an *N*-compound database or library (where, typically, $n << N$), which involves generating each of the

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

possible subsets and then calculating their diversities using a *diversity index* (as discussed below). Such a procedure is computationally infeasible for large values of *n* and *N* (Kuo *et al*., 1993) and there has thus been much interest in alternative approaches for selecting diverse sets of molecules, with three principal methods having been described thus far: *cluster-based* selection, *partition-based* selection and *dissimilarity-based* selection (Dean and Lewis, 1999).

Cluster analysis, or clustering, is the process of sub-dividing 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 (Everitt, 1993; Sneath and Sokal, 1973). It is thus possible 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 (Willett, 1987). Cluster-based methods have been widely used for molecular diversity studies (Brown and Martin, 1996, 1997; Dunbar, 1997; Shemetulskis *et al*., 1995) but they are increasingly being supplanted by dissimilarity-based and partition-based approaches.

Partition-based compound selection requires the identification of a set of *p* characteristics, these typically being molecular properties that would be expected to affect the ability of a small molecule to bind to a protein (Mason and Pickett, 1997). The range of values for each such characteristic is divided into a set of sub-ranges. The combinatorial product of all possible sub-ranges then defines a *p*-dimensional grid of bins (or cells) that is referred to as a partition, and each molecule is assigned to the bin that matches that molecule's set of characteristics. A subset is obtained by selecting one (or some small number) of the molecules from each of the bins. Partition-based selection is very fast in operation, and has the advantage that it permits the rapid identification of those sections of structural space that are under-represented, or even unrepresented, in a database (Pearlman and Smith, 1998).

Cluster-based and partition-based approaches identify a set of dissimilar molecules indirectly, since the approaches involve the identification of clusters or bins of similar molecules. Dissimilarity-based approaches, conversely, try to identify a set of dissimilar molecules in a

dataset directly, using some quantitative measure of dissimilarity (Lajiness, 1997). This class of approaches is discussed in detail in subsequent sections of this paper.

The many selection techniques that are available has encouraged interest in comparative studies to ascertain which are the most effective. Such studies are vitally important if one wishes to identify the best procedures, but they require some quantitative measure of effectiveness. This has led to the development of several diversity indices, which provide a single-number quantification of the degree of structural variation within a dataset. Examples of such approaches include a count of the number of bits that are set in the union of all of the fingerprints for a dataset (Martin *et al*., 1995), the number of distinct substructures that can be generated from all of the molecules in a dataset (Bone and Villar, 1997), the fraction of the bins in a partition that contain some minimal number of molecules (Pickett *et al*., 1996), and the sum of the pairwise inter-molecular dissimilarities for a dataset (Turner *et al*., 1997).

Having provided a brief overview of the current status of computational tools for the analysis of molecular diversity, we now focus on dissimilarity-based methods for compound selection, illustrating the range of procedures that are available by reference to work carried out over the last three years in the University of Sheffield (Gardiner *et al*., 1998; Gillet *et al*., 1997, 1999; Holliday *et al*., 1995; Snarey *et al*., 1998).

2. SELECTION OF COMPOUNDS FROM A DATABASE

The most obvious selection task is to identify a diverse subset of an entire database, this being most commonly done when there is a need to select some representative number of compounds from a company's corporate database for testing in a novel bioassay; this requirement, indeed, provided the rationale for the very first work on systematic methods for compound selection, using cluster-based approaches, back in the mid-Eighties (Willett, 1987). We have already noted that the identification of the n most diverse molecules in a dataset containing N molecules is generally infeasible for non-trivial values of n and N (but see Section 4 below for an exception to this general rule), and practicable approaches to dissimilarity-based compound selection hence involve approximate methods that are not guaranteed to result in the identification of the most dissimilar possible subset (see, *e.g*., Bawden, 1993; Clark, 1997; Hudson *et al*., 1996; Lajiness, 1990, Marengo and Todeschini, 1992; Nilakantan *et al*., 1997; Pickett *et al*., 1998; Polinsky *et al*., 1996); that said, there is evidence to suggest that the subsets identified are only marginally sub-

optimal (Gillet *et al*., 1997). Thus far, two major classes of algorithm have been described: *maximum-dissimilarity* algorithms and *sphere-exclusion* algorithms (Snarey *et al*., 1998)

The basic maximum-dissimilarity algorithm for selecting a size-*n* ***Subset*** from a size-*N* ***Dataset*** is shown in Figure 1. This algorithm, which was first described by Kennard and Stone (1969) and which was applied to compound selection by Lajiness (1990) and Bawden (1993), permits many variants depending upon the precise implementation of Steps 1 and 3. Possible mechanisms for the choice of the initial compound in Step 1 include: choosing a compound at random; choosing that compound that is most dissimilar to the other compounds in ***Dataset***; or choosing that compound that is nearest to the centre (in some sense) of ***Dataset***, *inter alia*. Step 3 in the figure requires a quantitative definition of the dissimilarity between a single compound in ***Dataset*** and the group of compounds that comprise ***Subset***, so that the most dissimilar molecule can be identified in each iteration of the algorithm.

There are several ways in which "most dissimilar" can be defined, with each definition resulting in a different version of the algorithm and hence in the selection of a different subset (Holliday and Willett, 1996) (in just the same way as different clustering methods result from the use of different similarity criteria in hierarchic agglomerative clustering (Lance and Williams, 1967). Examples of such definitions include MaxSum (Pickett *et al*., 1998) and MaxMin (Polinsky *et al*., 1996). Let ***DIS***(***A***,***B***) be the dissimilarity between two molecules, or sets of molecules, ***A*** and ***B***. Consider a single compound, ***J***, taken from ***Dataset*** and the ***m*** compounds that form the current membership of ***Subset*** at some stage in the selection process; then the dissimilarity between ***J*** and ***Subset***, ***DIS***(***J***, ***Subset***), is given by

$$\sum DIS(J,K) \text{ and } minimum\{DIS(J,K)\}$$

in the case of the MaxSum and MaxMin definitions, respectively, with K ($1 \leq K \leq m$) ranging over all of the ***m*** molecules in ***Subset*** at that point. The molecule chosen for addition to ***Subset*** is then that with the largest value of ***DIS***(***J***, ***Subset***).

Insert Figure 1 about here

The basic maximum dissimilarity algorithm shown in Figure 1 has an expected time complexity of $O(n^2N)$; as n is normally some small fraction of N (such as 1% or 5%), this represents a running time that is cubic in N, which makes it extremely demanding of computational resources if ***Dataset*** is at all large. Holliday *et al*. (1995) described a MaxSum selection algorithm with a time

complexity of $O(nN)$, using an equivalence that had been developed for the rapid implementation of hierarchic agglomerative document clustering using the group-average clustering method (Voorhees, 1986). However, an analysis of the MaxSum definition by Agrafiotis and Lobanov (1999) suggested that it could result in subsets containing groups of closely-related molecules, and this limitation was subsequently demonstrated by Snarey *et al.* (1998) in a comparison of several different methods for dissimilarity-based compound selection. Although not as fast in practice as MaxSum, MaxMin can also be implemented with an $O(nN)$ algorithm (Higgs *et al.*, 1997; Polinsky *et al.*, 1996) and the comparative evaluation of Snarey *et al.* (1998) showed it to be more effective than MaxSum in identifying database subsets exhibiting a range of biological activities; accordingly, it is probably the method of choice for this class of selection algorithms.

A further variant of the basic approach shown in Figure 1 is to specify a threshold dissimilarity, *t*, and then to reject the molecule selected in Step 2 if it has a dissimilarity less than *t* with any of the compounds already in *Subset*. The inclusion of such a threshold results in a maximum dissimilarity algorithm that is not too far removed from the basic sphere-exclusion approach described by Hudson *et al.* (1996). Here, a threshold *t* is set, which can be thought of as the radius of a hypersphere in multi-dimensional chemistry space. A compound is selected, either at random or using some rational basis, for inclusion in *Subset* and the algorithm then excludes from further consideration all those other compounds within the sphere centred on that selected compound, as shown in Figure 2. Many variants are again possible, depending upon the manner in which Stage 2 is implemented. Thus, one can choose that molecule that is most dissimilar to the existing *Subset*, in which case different results will be obtained (as with the maximum dissimilarity algorithms) depending upon the dissimilarity definition that is adopted. Alternatively, a compound can be selected at random, as in the MDISS (DiverseSolutions, 1996) and DIVPIK (Nilakantan *et al.*, 1997) programs, this resulting in an exceptionally fast, but non-deterministic, algorithm. Several examples of sphere-exclusion algorithms were evaluated by Snarey *et al.* (1998), who found that they were broadly comparable in performance to the MaxMin maximum dissimilarity algorithm.

Insert Figure 2 about here

The close relationship that exists between these two classes of algorithm has recently been highlighted by Clark (1997), who describes a program called OptiSim (for Optimizable K-Dissimilarity Selection) that is summarised in Figure 3 and that makes use of an intermediate pool

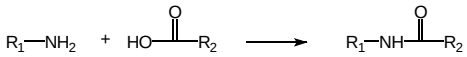
of selected compounds, here called ***Subsample***. An inspection of this figure shows that the mode of processing is determined by the value of ***K*** (the size of ***Subsample***) that is specified, with values of ***K*** equal to 1 and to ***N*** corresponding to (versions of) sphere-exclusion and maximum dissimilarity, respectively. Clark presents a detailed discussion of how the choice of ***K*** affects the behaviour of the algorithm and the trade-offs that are to be expected between what he describes as the ***representativeness*** of subsets generated by sphere-exclusion methods and the ***diversity*** of subsets generated by maximum-dissimilarity methods; a further discussion of these characteristics is provided by Clark and Langton (1998).

Insert Figure 3 about here

In concluding this section, it is perhaps worth noting that the use of any selection procedure should be preceded by the application of a filtering mechanism to ensure the removal from further consideration of those molecules that exhibit some sort of undesirable characteristic (Walters ***et al***., 1998). Examples of such characteristics include: the presence in a molecule of highly reactive or toxic substructures that have been catalogued in a corporate "badlist" of undesirable fragments (Lajiness, 1997); and restrictions on the values of properties such as the molecular weight, the octanol-water partition coefficient, and the numbers of rotatable bonds and chiral centres (Lipinski ***et al***., 1997). Similar comments apply to the library design procedures discussed in the following sections of this paper.

3. REAGENT-BASED DESIGN OF COMBINATORIAL LIBRARIES

Thus far, we have considered dissimilarity-based algorithms for identifying a subset of an entire database; however, the rapid development of combinatorial chemistry methods has spurred the development of selection tools for designing combinatorial libraries. Consider the simple amide reaction shown below, in which a primary amine is coupled with a carboxylic acid:



Databases of commercially available compounds, such as ***the Available Chemicals Directory***, will reveal literally thousands of amines and acids (the R_1 and R_2, respectively) that might be used for this reaction: it is thus possible, in principle at least, to create a combinatorial library containing millions of amides if all of the possible reactions were to be carried out. The similar property principle (Johnson and Maggiora, 1990) states that structurally similar molecules are expected to

exhibit similar properties and activities; thus, if many of the products in the reaction above are similar to each other, there is little novel structure-activity information to be gained from their synthesis. Accordingly, it should be possible to increase the cost-effectiveness of lead-discovery programmes by synthesising just a subset of the entire combinatorial library, subject to that subset encompassing the full range of structural types present in that library. One normally differentiates between a combinatorial library, which is the one that is actually synthesised and tested, and a *virtual library*, which is the library that would be obtained by the exhaustive enumeration of all of the possible products. Typically, a virtual library exists only as a computer data structure, from which the actual combinatorial library is chosen with some selection algorithm (Cramer *et al*., 1998).

There are two basic approaches to the design of structurally diverse combinatorial libraries. The initial approach, first described in detail by Martin *et al*. (1995), takes as its basis the assumption that if it is possible to identify maximally diverse (or, more realistically, near maximally diverse) sets of reactants, then their use will result in the generation of a maximally diverse combinatorial library of products when the reactants are combined in a combinatorial synthesis. This *reagent-based* approach is computationally attractive, as it means that the selection algorithm need only be applied to the individual sets of reagents, and it rapidly established itself as the method of choice for designing combinatorial libraries. For example, assume that there are 1000 acids and 1000 amines available within a company's corporate files, and that there is the capacity to synthesise and test 10,000 amides; then these amides can be achieved by selecting 100 structurally diverse acids and 100 structurally diverse amines. Analysis of the full virtual library, conversely, would require consideration of all 1,000,000 possible amide products, and this second, *product-based* approach thus received little attention until recently (as discussed further in Section 4 of this paper).

Any of the selection algorithms described above can be used for reagent-based selection, with the need for structural diversity increasingly being complemented by consideration of the physicochemical properties of the molecules involved (Martin and Critchlow, 1999). Here, we describe a maximum dissimilarity selection algorithm that has been designed for reagent-based selection and that has the ability to find all possible subsets that satisfy an external diversity criterion, rather than just the single subset that is the output of most other algorithms that have been designed for this purpose. The starting point for the work was a detailed verification of the similar property principle carried out by Brown and Martin (1996, 1997). Given a molecule, *I*, of

known activity, Brown and Martin show that there is a high *a priori* probability that any near neighbour of *I* will also be active, where a near neighbour is deemed to be one that has a Tanimoto similarity of at least 0.85 when *I* and each of the other molecules in a database are characterised by Tripos UNITY 2D fingerprints. While it is most unlikely that this precise value provides the best cut-off for all possible types of biological activity, it does provide a very simple basis for dissimilarity-based compound-selection, by applying a (dis)similarity threshold in Step 3 of Figure 1 to ensure that no two molecules in *Subset* will be strongly similar to each other. The algorithm of Gardiner *et al*. (1998) is designed to identify all such subsets that satisfy this dissimilarity criterion.

Let *M* be an $N \times N$ dissimilarity matrix in which $M(I,J)$ contains the dissimilarity between the *I*-th and *J*-th compounds in a dataset containing *N* compounds (typically this dataset will be all of the available reagents of some particular type, *e.g*., all of the primary amines in the *Available Chemicals Directory* or in a company's corporate database). A *subset-selection graph*, *G*, is created from *M* by applying a threshold dissimilarity, *t*, and then setting each element $M(I,J)$ to one (or zero) depending upon whether it is greater than (or not greater than) the threshold. The complete set of subsets satisfying the dissimilarity criterion is then the set of *cliques* of size *n* (*i.e*., containing *n* vertices) in *G*, where a clique is a subgraph in which every vertex is connected to every other vertex and which is not contained in any larger subgraph with this property. Clique detection is known to be NP-complete, except in the case of special types of graph, and the observation that diverse-subset selection and clique detection are equivalent is of little practical use unless it is possible to identify clique-detection algorithms that are sufficiently rapid in execution to permit the processing of subset-selection graphs of non-trivial size. Algorithms for clique-detection in graphs have been extensively studied (Pardalos and Xue, 1994). Gardiner *et al*. report a comparison of several such algorithms when applied to the processing of subset-selection graphs, and suggest that one due to Babel (1991) is sufficiently fast to enable the procedure to be applied to the selection of reagents for combinatorial synthesis. Once all of the subsets have been generated by the procedure, which is summarised in Figure 4, a further filtering step (based on criteria such as cost, physicochemical-parameter or diversity-index values or other characteristics such as those discussed at the end of Section 2), can be employed to identify the particular subset that will be chosen for use in some application (Gardiner *et al*., 1998).

Insert Figure 4 about here

4. PRODUCT-BASED DESIGN OF COMBINATORIAL LIBRARIES

Consider a combinatorial library, *c*, that is synthesised from reactants contained in two *reactant pools*, r_1 and r_2, of sizes n_1 and n_2, respectively (in the following, we consider only dimer libraries for the purpose of simplicity but the analysis can be extended to reactions that involve a greater number of reactants). These two reactant pools have previously been selected as representing diverse subsets of two larger *potential-reactant pools*, R_1 and R_2, of sizes N_1 and N_2, respectively, using some subset-selection algorithm. Let *V* be the corresponding virtual library, *i.e.*, the fully enumerated combinatorial library that would have been generated from all possible combinations of R_1 and R_2 if the subset-selection procedure had not been used. Thus, *c* and *V* contain n_1n_2 and N_1N_2 dimers, respectively. The assumption underlying reagent-based selection is that the library *c* will be as diverse as a library obtained by employing the same subset-selection procedure that was used to create the reactant pools r_1 and r_2, (*i.e.*, that was used to identify the n_1 most dissimilar molecules in R_1 and the n_2 most dissimilar molecules in R_2) to identify the most dissimilar n_1n_2-molecule library from amongst the N_1N_2 molecules in *V*. This subset is referred to subsequently as library *L*.

The validity of this assumption was challenged by Gillet *et al*. (1997), who took three published combinatorial syntheses, generated libraries by both of the procedures described above, and then calculated the diversities of the two libraries using the diversity index described by Turner *et al*. (1997): in all cases, the library *L* had a diversity that was greater than that of the library *c*. Thus, the greater effort involved in generating *L*, which involves the analysis of $N_1{\times}N_2$ product molecules as against the analysis of the $N_1{}^{+}N_2$ reactant molecules required to generate *c*, results in an increase in the diversity of the final library. However, while *L* is a library, it is not a combinatorial library in that it contains a maximally diverse set of independent product molecules, rather than a set that can be synthesised using a combinatorial reaction.

The synthetic inefficiency that can result from performing selection at the product level is illustrated in Figure 5(a), in which a virtual library, *V*, built from two reactant pools is represented by a 9×9 matrix. The rows of the matrix represent the N_1 reactants (x_1...x_9) available in pool R_1, and the columns of the matrix represent the N_2 reactants (y_1...y_9) in pool R_2. The N_1N_2 elements of the matrix then represent the full combinatorial library, *V*, that would result from reacting all

the reactants in R_1 with all the reactants in R_2. Assume that we wish L to contain the nine most diverse compounds from C. Then a selection algorithm can select compounds from anywhere within the matrix: for example, the resulting library might correspond to the shaded elements, as shown: such a "cherrypicking" approach is, of course, analogous to that employed when selecting individual compounds from a full database, as discussed in Section 3. The potential synthetic inefficiency of this approach is highlighted by the fact that thirteen reactants are required to build the nine-member library (*viz* six reactants - x_3, x_4, x_5, x_6, x_7 and x_8 - from pool R_1 and seven reactants - y_1, y_2, y_4, y_6, y_7, y_8 and y_9 - from pool R_2), rather than the three from each pool required to build a nine-member subset that is a combinatorial library.

A nine-member subset of V that does represent a true combinatorial library can be selected by intersecting three rows of the matrix with three columns: for example, a 3×3 library built from reactants x_3, x_6 and x_8 reacted with reactants y_2, y_4 and y_5 is shown by the shaded elements of the matrix in Figure 5(b). Finding the optimal library then requires consideration of all possible n_1n_2-member sets of products obtained by reacting combinations of n_1 reagents selected from R_1 and n_2 reagents selected from R_2. We have devised a genetic algorithm (GA) for this demanding search problem, with each chromosome in the population representing one possible combinatorial library. For an n_1n_2-member library, a chromosome consists of two parts: the first part represents the n_1 reactants selected from pool R_1 (*i.e.*, the rows of the matrix) and the second part represents the n_2 reactants selected from pool R_2 (*i.e.*, the columns of the matrix). The fitness of a chromosome is obtained by constructing the n_1n_2-member combinatorial library represented by it, and then calculating the diversity of this library using a diversity index. The index used here was the mean pairwise dissimilarity (specifically the complement of the Tanimoto coefficient) when averaged over all the pairs of molecules in a size-n_1n_2 library, the molecules being represented by molecular fingerprints. This index is discussed by Pickett *et al.* (1998) and Turner *et al.* (1997) and was used here since it can be calculated very rapidly, a pre-requisite for use in a GA-based application where very large numbers of fitness values may need to be calculated. The GA operators are applied to maximise the average diversity and hence to identify the maximally diverse library.

Insert Figure 5 about here

Experiments with several published combinatorial library designs showed that the diversities of the libraries resulting from the GA's product-based selection were consistently greater than the diversities of the corresponding libraries resulting from conventional reagent-based selection (Gillet *et al*., 1997). As well as being effective in operation, the algorithm is also surprisingly efficient given the size of the search-space that needs to be explored, with the selection of 40×40 reagent pools from a 160,000-member virtual library requiring approximately 20 minutes for a C program running on a Silicon Graphics R10000 processor. In addition to choosing a structurally diverse combinatorial library, the SELECT program of Gillet *et al*. (1999) also ensures that the constituent molecules exhibit "drug-like" properties. This is achieved by means of a multi-objective fitness function of the form

$$w_D(D) + w_C(C) + w_{f1}\Delta f1 + w_{f2}\Delta f2\ldots$$

where the first term describes the diversity of the library that is being designed, as in the basic version of the GA described previously. The second term is designed to force the library to be different from some existing reference collection; for example, it may be desirable to ensure that the library is maximally dissimilar from a library that has already been synthesised and tested. The remaining terms in the fitness function relate to physical properties of molecules that are thought to affect their ability to function as a drug (such as the molecular weight, the numbers of rotatable bonds, hydrogen donors and acceptors, and the octanol/water partition coefficient) and that can be calculated sufficiently rapid for the processing of libraries of realistic size. A physical property of the library is optimised by comparing the distribution of its values in the library with the distribution of values of the same property in some reference collection (for which we use the *World Drugs Index* database of known drugs). The various *w* terms act as weights that reflect the relative importance of each of the various components of the fitness function, thus allowing the designer to control the sorts of library that are produced (Gillet *et al.*, 1999).

Other types of combinatorial search algorithm can, of course, be used to explore library space, and there have been several reports of the use of simulated annealing (SA) for library design. In this work, molecules are represented by principal components derived from calculated physical properties (topological and information content indices, and electronic, hydrophobic and steric descriptors) (Hassan *et al*., 1996) or by low-dimensionality autocorrelation vectors describing the distribution of the electrostatic potential over the van der Waals' surface of a molecule (Agrafiotis, 1997), and the scoring function for the SA uses one of several different inter-molecular distance functions in the resulting descriptor space. Another example of the use of SA as a searching tool is provided by the HARPick program (Good and Lewis, 1997). Here, a molecule is characterised

by its constituent three-point pharmacophores, these being generated from an approximate 3D structure, and the diversity of a set of molecules, such as a putative combinatorial library, is given by a function based on the number of distinct pharmacophores present in that particular library. As with SELECT, the scoring function encompasses not just structural diversity but also a physicochemical property. Specifically, an attempt is made to ensure an approximately even distribution of a library's members across three properties that provide a crude, but rapidly computable, measure of molecular shape: these are the number of heavy atoms in a molecule, the largest triangle perimeter for any of the three-point pharmacophores in that molecule, and the largest triangle area for any of these pharmacophores.

5. CONCLUSIONS

Computational methods for the processing of chemical structure information have been used to support drug-discovery programmes for more than three decades; however, the introduction of combinatorial approaches to drug discovery has now focused the industry's interest on tools that can analyse and control the floods of chemical and biological data resulting from such approaches. In this paper, we have reviewed some of the techniques that have been developed for selecting diverse sets of compounds from chemical structure databases, both real (as with corporate structure files or files of available synthetic reagents) and virtual (as with fully enumerated combinatorial libraries), illustrating the techniques by focusing upon methods for dissimilarity-based selection that have been developed in our laboratory in Sheffield.

Although the techniques discussed here, and many others that have been reviewed elsewhere (Dean and Lewis, 1999), provide effective and efficient means of selecting compounds there is still much scope for further work. A very simple, but important, task is that of comparing the many available methods to find those that are most suitable, both in terms of efficiency and effectiveness. Such studies are now appearing in the literature (Brown and Martin, 1996, 1997; Matter, 1997; Patterson *et al*., 1996) but only that by Snarey *et al*. (1998) has provided a detailed analysis of some of the dissimilarity-based selection methods considered in this paper. More importantly, the full value of methods for analysis molecular diversity will only be obtained when they are linked to other, existing approaches to computer-aided molecular design, such as ligand docking, pharmacophore mapping and quantitative structure-activity relationships (Martin and Willett, 1998): the merits of such linked approaches are well illustrated by very recent work on the

docking of combinatorial libraries (Jones *et al*., 1999; Kick *et al*., 1997), and we can expect many further such reports in the next few years.

ACKNOWLEDGEMENTS

I thank the following: John Bradshaw, Eleanor Gardiner, Val Gillet, Darren Green, John Holliday, Mike Snarey, Nick Terret and David Wilton for their contributions to the work reported here; the Engineering and Physical Sciences Research Council, GlaxoWellcome, Pfizer Central Research and Tripos Inc. for funding; and Daylight Chemical Information Systems and Tripos Inc. for software support. The Krebs Institute for Biomolecular Research is a designated centre of the Biotechnology and Biological Sciences Research Council.

REFERENCES

Agrafiotis, D.K. 1997. Stochastic algorithms for maximising molecular diversity. *J. Chem. Inf. Comput. Sci*. 37, 841-851.

Agrafiotis, D. and Lobanov, V.S. 1999. An efficient implementation of distance-based diversity measures based on *k-d* trees. *J. Chem. Inf. Comput. Sci*. 39, 51-58.

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

Babel, L. 1991. Finding maximum cliques in arbitrary and special graphs. *Computing* 46, 321-341.

Bawden, D. 1993. Molecular dissimilarity in chemical information systems, 383-388. *In* Warr, W.A., ed., *Chemical Structures 2. The International Language of Chemistry*. Springer-Verlag, Heidelberg.

Bone, R.G.A. and Villar, H.O. 1997. Exhaustive enumeration of molecular substructures. *J. Comput. Chem.* 18, 86-107.

Brown, R.D. 1997. Descriptors for diversity analysis. *Perspect. Drug. Disc. Design* 7/8, 31-49.

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. *J. Chem. Inf. Comput. Sci*. 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. *J. Chem. Inf. Comput. Sci.* 37, 1-9.

Chaiken, I.M. and Janda, K.D., eds., 1996. *Molecular Diversity and Combinatorial Chemistry. Libraries and Drug Discovery*, American Chemical Society, Washington.

Clark, R.D. 1997. OptiSim: an extended dissimilarity selection method for finding diverse representative subsets. *J. Chem. Inf. Comput. Sci*. 37, 1181-188.

Clark, R.D. and Langton, W.J. 1998. Balancing representativeness against diversity using optimizable *k*-dissimilarity and hierarchical clustering. *J. Chem. Inf. Comput. Sci*. 38, 1079-1086,

Cramer, R.D., Patterson, D.E., Clark, R.D., Soltanshahi, F. and Lawless, M.S. 1998. Virtual compound libraries: a new approach to decision making in molecular discovery research. *J. Chem. Inf. Comput. Sci*. 38, 1010-1023.

Dean, P.M. and Lewis, R.A., eds., 1999. *Molecular Diversity in Drug Design*, in the press.

DeWitt, S.H. and Czarnik, A.W., eds., 1997. *A Practical Guide to Combinatorial Chemistry*, American Chemical Society, Washington.

DiverseSolutions User's Manual 1996. Tripos Inc., St Louis.

Downs, G.M. and Willett, P. 1995. Similarity searching in databases of chemical structures. *Rev. Comput. Chem.* 7, 1-66.

Dunbar, J.B. 1997. Cluster-based selection. *Perspect. Drug. Disc. Design* 7/8, 51-63.

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

Gardiner, E.J., Artymiuk, P.J. and Willett, P. 1998. Clique-detection algorithms for matching three-dimensional molecular structures. *J. Mol. Graph. Model.* 15, 245-253.

Gillet, V.J., Willett, P. and Bradshaw, J. 1997. The effectiveness of reactant pools for generating structurally diverse combinatorial libraries. *J. Chem. Inf. Comput. Sci.* 37, 731-740.

Gillet, V.J., Willett, P., Bradshaw, J. and Green, D.V.S. 1999. Selecting combinatorial libraries to optimise diversity and physical properties. *J. Chem. Inf. Comput. Sci*. 39, 169-177.

Good, A.C. and Lewis, R.A. 1997. New methodology for profiling combinatorial libraries and screening sets: cleaning up the design process with HARPick. *J. Med. Chem.* 40, 3926-3936.

Hassan, M., Bielawski, J.P., Hempel, J.C. and Waldman, M. 1996. Optimization and visualization of molecular diversity of combinatorial libraries. *J. Comput.-Aid. Mol. Design* 2, 64-74.

Higgs, R.E., Bemis, K.G., Watson, I.A. and Wikel, J.H. 1997. Experimental designs for selecting molecules from large chemical databases. *J. Chem. Inf. Comput. Sci.* 37, 861-870.

Holliday, J.D., Ranade, S.S. and Willett, P. 1995. A fast algorithm for selecting sets of dissimilar structures from large chemical databases. *Quant. Struct.-Activ. Relat.* 14, 501-506.

Holliday, J.D. and Willett, P. 1996. Definitions of 'dissimilarity' for dissimilarity-based compound selection. *J. Biomolec. Screen.* 1, 145-151.

Hudson, B.D., Hyde, R.M., Rahr, E. and Wood, J. 1996. Parameter based methods for compound selection from chemical databases. *Quant. Struct.-Activ. Relat.* 15, 285-289.

Johnson, M.A. and Maggiora, G.M., eds., 1990. *Concepts and Applications of Molecular Similarity*, John Wiley, New York.

Jones, G., Willett, P., Glen, R.C., Leach, A.R. and Taylor, R. 1999. Further development of a genetic algorithm for ligand docking and its application to screening combinatorial libraries. In the press.

Kennard, R.W. and Stone, L.A. 1969. Computer aided design of experiments. *Technometrics* 11, 137-148.

Kick, E.K., Roe, D.C., Skillman, A.G., Liu, G., Ewing, T.J.A., Sun, Y., Kuntz, I.D. and Ellman, J.A. 1997. Structure-based design and combinatorial chemistry yield low nanomolar inhibitors of cathepsin D. *Chem. Biol*. 4, 297-307.

Kuo, C.-C., Glover, F. and Dhir, K.S. 1994. Analyzing and modeling the maximum diversity problem by zero-one programming. *Decision Sci*. 24, 1171-1185.

Lajiness, M.S. 1990. Molecular similarity-based methods for selecting compounds for screening, 299-316. *In* Rouvray, D.H., ed., *Computational Chemical Graph Theory*, Nova Science Publishers, New York.

Lajiness, M.S. 1997. Dissimilarity-based compound selection techniques. *Perspect. Drug. Disc. Design* 7/8, 65-84.

Lance, G.N. and Williams, W.T. 1967. A general theory of classificatory sorting strategies. I. Hierarchical systems. *Comput. J.* 9, 373-380.

Lipinski, C.A., Lombardo, F., Dominy, B.W. and Feeney, P.J. 1997. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Res.*, 23, 3-25.

Marengo, E. and Todeschini, R. 1992. A new algorithm for optimal, distance-based experimental design. *Chemomet. Intell. Lab. Syst.* 16, 37-44.

Martin, E.J., Blaney, J.M., Siani, M.A., Spellmeyer, D.C., Wong, A.K. and Moos, W.H. 1995. Measuring diversity: experimental design of combinatorial libraries for drug discovery. *J. Med. Chem*. 38, 1431-1436.

Martin, E.J. and Critchlow, R.E. 1999. Beyond mere diversity: tailoring combinatorial libraries for drug discovery. *J. Combin. Chem.* 1, 32-45.

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

Mason, J.S. and Pickett, S.D. 1997. Partition-based selection. *Perspect. Drug. Disc. Design* 7/8, 85-114.

Matter, H. 1997. Selecting optimally diverse compounds from structure databases: a validation study of two-dimensional and three-dimensional molecular descriptors. *J. Med. Chem.* 40, 1219-1229.

Nilakantan, R., Bauman, N. and Haraki, K.S. 1997. Database diversity assessment: new ideas, concepts and tools. *J. Comput.-Aid. Mol. Design* 11, 447-452.

Pardalos, P.M. and Xue, J. 1994. The maximum clique problem. *J. Global Optimiz.* 4, 301-328.

Patterson, D.E., Cramer, R.D., Ferguson, A.M., Clark, R.D. and Weinberger, L.E. 1996. Neighbourhood behaviour: a useful concept for validation of "molecular diversity" descriptors. *J. Med. Chem.* 39, 3049-3059.

Pearlman, R.S. and Smith, K.M. 1998. Novel software tools for chemical diversity. *Perspect. Drug Disc. Design* 9/11, 339-353.

Pickett, S.D., Luttman, C., Guerin, V., Laoui, A. and James, E. 1998. DIVSEL and COMPLIB - strategies for the design and comparison of combinatorial libraries using pharmacophore descriptors. *J. Chem. Inf. Comput. Sci.* 38, 144-150.

Pickett, S.D., Mason, J.S. and McLay, I.M. 1996. Diversity profiling and design using 3D pharmacophores: pharmacophore-derived queries (PDQ). *J. Chem. Inf. Comput. Sci*. 36, 1214-1223.

Polinsky, A., Feinstein, R.D., Shi, S. and Kuki, A. 1996. LiBrain: software for automated design of exploratory and targeted combinatorial libraries, 219-232. *In* Chaiken, I.M. and Janda, K.D., eds., *Molecular Diversity and Combinatorial Chemistry. Libraries and Drug Discovery*, American Chemical Society, Washington DC.

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. *J. Comput.-Aid. Mol. Design* 9, 407-416.

Snarey, M., Terret, N.K., Willett, P. and Wilton, D.J. 1998. Comparison of algorithms for dissimilarity-based compound selection. *J. Mol. Graph. Model*. 15, 372-385.

Sneath, P.H.A. and Sokal, R.R. 1973. *Numerical Taxonomy*, WH Freeman, San Francisco.

Turner, D.B., Tyrrell, S.M. and Willett, P. 1997. Rapid quantification of molecular diversity for selective database acquisition. *J. Chem. Inf. Comput. Sci.* 37, 18-22.

Voorhees, E.M. 1986. Implementing agglomerative hierarchic clustering algorithms for use in document retrieval. *Inf. Proc. Manag*. 22, 465-476.

Walters, W.P., Stahl, M.T. and Murcko, M.A. 1998. Virtual screening - an overview. *Drug Discov. Today*, 3, 160-178.

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

Willett, P., Barnard, J.M. and Downs, G.M. 1998. Chemical similarity searching. *J. Chem. Inf. Comput. Sci*. 38, 983-996.

1. Initialise *Subset* by transferring a compound from *Dataset*.
2. Calculate the dissimilarity between each remaining compound in *Dataset* and the compounds in *Subset*.
3. Transfer to *Subset* that compound from *Dataset* that is most dissimilar to *Subset*.
4. Return to Step 2 if there are less than *n* compounds in *Subset*.

Fig. 1. General maximum-dissimilarity algorithm

1. Define a threshold dissimilarity, *t*.
2. Transfer a compound, *J*, from *Dataset* to *Subset*.
3. Remove from *Dataset* all compounds having a dissimilarity with *J* of less than *t*.
4. Return to Step 2 if there are compounds remaining in *Dataset*.

Fig. 2. General sphere-exclusion algorithm

1. Define a threshold dissimilarity, *t*.
2. Initialise *Subset* by transferring a compound, *J*, from *Dataset*.
3. Select a compound, *J*, from *Dataset*. If it has a dissimilarity less than *t* with any compound in *Subset* then remove it from *Dataset*; otherwise add it to *Subsample*.
4. Repeat Step 3 until *Subsample* contains *K* molecules.
5. Transfer to *Subset* that compound from *Subsample* that is most dissimilar to *Subset*. Return the remaining members of *Subsample* to *Dataset*.
6. Return to Step 3 if there are less than *n* compounds in *Subset*.

Fig. 3. OptiSim algorithm (Clark, 1997).

1. Define a threshold dissimilarity, *t*.
2. Generate an $N \times N$ dissimilarity matrix in which $M(I,J)$ contains the dissimilarity between molecules *I* and *J*.
3. Generate a graph, *G*, from *M* by setting each element $M(I,J)$ to one (or zero) if it is greater than (or not greater than) *t*.
4. Use a clique-detection algorithm to identify the set of size-*n* cliques in *G*.

Fig. 4. Clique-based processing to identify all subsets meeting a dissimilarity criterion

	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9
x_1	$x_1 y_1$	$x_1 y_2$	$x_1 y_3$	$x_1 y_4$	$x_1 y_5$	$x_1 y_6$	$x_1 y_7$	$x_1 y_8$	$x_1 y_9$
x_2	$x_2 y_1$	$x_2 y_2$	$x_2 y_3$	$x_2 y_4$	$x_2 y_5$	$x_2 y_6$	$x_2 y_7$	$x_2 y_8$	$x_2 y_9$
x_3	$x_3 y_1$	$x_3 y_2$	$x_3 y_3$	$x_3 y_4$	$x_3 y_5$	$x_3 y_6$	$x_3 y_7$	$x_3 y_8$	$x_3 y_9$
x_4	$x_4 y_1$	$x_4 y_2$	$x_4 y_3$	$x_4 y_4$	$x_4 y_5$	$x_4 y_6$	$x_4 y_7$	$x_4 y_8$	$x_4 y_9$
x_5	$x_5 y_1$	$x_5 y_2$	$x_5 y_3$	$x_5 y_4$	$x_5 y_5$	$x_5 y_6$	$x_5 y_7$	$x_5 y_8$	$x_5 y_9$
x_6	$x_6 y_1$	$x_6 y_2$	$x_6 y_3$	$x_6 y_4$	$x_6 y_5$	$x_6 y_6$	$x_6 y_7$	$x_6 y_8$	$x_6 y_9$
x_7	$x_7 y_1$	$x_7 y_2$	$x_7 y_3$	$x_7 y_4$	$x_7 y_5$	$x_7 y_6$	$x_7 y_7$	$x_7 y_8$	$x_7 y_9$
x_8	$x_8 y_1$	$x_8 y_2$	$x_8 y_3$	$x_8 y_4$	$x_8 y_5$	$x_8 y_6$	$x_8 y_7$	$x_8 y_8$	$x_8 y_9$
x_9	$x_9 y_1$	$x_9 y_2$	$x_9 y_3$	$x_9 y_4$	$x_9 y_5$	$x_9 y_6$	$x_9 y_7$	$x_9 y_8$	$x_9 y_9$

(a)

	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9
x_1	$x_1 y_1$	$x_1 y_2$	$x_1 y_3$	$x_1 y_4$	$x_1 y_5$	$x_1 y_6$	$x_1 y_7$	$x_1 y_8$	$x_1 y_9$
x_2	$x_2 y_1$	$x_2 y_2$	$x_2 y_3$	$x_2 y_4$	$x_2 y_5$	$x_2 y_6$	$x_2 y_7$	$x_2 y_8$	$x_2 y_9$
x_3	$x_3 y_1$	$x_3 y_2$	$x_3 y_3$	$x_3 y_4$	$x_3 y_5$	$x_3 y_6$	$x_3 y_7$	$x_3 y_8$	$x_3 y_9$
x_4	$x_4 y_1$	$x_4 y_2$	$x_4 y_3$	$x_4 y_4$	$x_4 y_5$	$x_4 y_6$	$x_4 y_7$	$x_4 y_8$	$x_4 y_9$
x_5	$x_5 y_1$	$x_5 y_2$	$x_5 y_3$	$x_5 y_4$	$x_5 y_5$	$x_5 y_6$	$x_5 y_7$	$x_5 y_8$	$x_5 y_9$
x_6	$x_6 y_1$	$x_6 y_2$	$x_6 y_3$	$x_6 y_4$	$x_6 y_5$	$x_6 y_6$	$x_6 y_7$	$x_6 y_8$	$x_6 y_9$
x_7	$x_7 y_1$	$x_7 y_2$	$x_7 y_3$	$x_7 y_4$	$x_7 y_5$	$x_7 y_6$	$x_7 y_7$	$x_7 y_8$	$x_7 y_9$
x_8	$x_8 y_1$	$x_8 y_2$	$x_8 y_3$	$x_8 y_4$	$x_8 y_5$	$x_8 y_6$	$x_8 y_7$	$x_8 y_8$	$x_8 y_9$
x_9	$x_9 y_1$	$x_9 y_2$	$x_9 y_3$	$x_9 y_4$	$x_9 y_5$	$x_9 y_6$	$x_9 y_7$	$x_9 y_8$	$x_9 y_9$

(b)

Fig. 5. A fully enumerated, dimer library (*V*) represented by a 9×9 matrix. In (a) the shaded elements represent an example of a subset library, *L*, that contains the nine most diverse

compounds and that was chosen by applying a selection algorithm to V. In (b) the $n_1 n_2$ subset of V is also a combinatorial library that can be selected by intersecting n_1 rows with n_2 columns.