Combinatorial Libraries And The Evaluation Of Diversity

Peter Willett

Introduction

The dramatic developments in combinatorial chemistry and high-throughput screening that have occurred over the last decade have enabled the pharmaceutical and agrochemical industries to move from an inherently sequential to an inherently parallel mode of processing, in which large numbers of compounds are synthesised and tested simultaneously. However, simple synthetic strategies can lead to the creation of closely-related molecules, which are likely to exhibit comparable bioactivity profiles and which thus provide little useful information to guide a discovery programme. How, then, can we minimise such redundant experimentation whilst still ensuring the identification of sufficient, structurally disparate molecules to permit the development of robust structure-activity relationships? *Molecular diversity analysis* is the name generally given to the computational techniques that are being developed to answer this question, with the term 'diversity' being taken to imply concepts such as 'different', 'dissimilar' and 'heterogeneous'; more specifically a diverse set of molecules is normally taken to be one that covers the largest possible expanse of chemical space (however defined) in the search for novel bioactive leads. There are many aspects to molecular diversity analysis [1-4]: here, we focus on just two, these being the ways in which diversity can be expressed in quantitative terms, and the ways in which libraries of compounds can be designed to be maximally diverse. We illustrate these aspects by reference to a program, called SELECT, that we have developed for the design of structurally diverse, druglike combinatorial libraries [5].

Quantification Of Diversity

Although widely used, 'diversity' is frequently expressed in qualitative terms, it needs to be defined in quantitative terms if we are to develop computer programs that seek to identify maximally diverse libraries. This is commonly (but not invariably) done by calculating the similarities between sets of molecules. There are many ways in which inter-molecular similarities can be calculated [6], with the two most important components of any similarity measure being the ***structural descriptors*** that are used to characterise the molecules, and the ***similarity coefficient*** that is used to quantify the degree of similarity between pairs of molecules.

Most molecular diversity studies have involved characterising molecules by sets of fragment substructures or sets of calculated physicochemical properties. The presence of fragment substructures (normally two-dimensional [2D] substructures containing patterns of connected atoms and bonds) within a molecule are encoded by setting bits in

a *bit-string* (or *fingerprint*). Examples of the sorts of fragment substructure that are used for this purpose are shown in Figure 1. Structure representations such as these have been used successfully in very many diversity studies, and there is also increasing interest in the use of three-dimensional (3D) substructural descriptors based upon sets of potential pharmacophoric patterns, such as H-bond donors and acceptors or ring centroids. Alternatively, physicochemical properties (normally calculated from a 2D molecular structure, but increasingly also from a 3D one generated by a program such as CONCORD or CORINA) can be used to describe topological, electronic, lipophilic or geometric features, with a molecule represented by a real-numbered vector that includes several such properties. Once molecules have been characterised in this way, the similarity between a pair of them is calculated by means of a similarity coefficient, which provides a numeric quantification of the degree of resemblance between two sets of such characterisations [7]. Similarity calculations employing substructural data (both 2D and 3D) have generally used association coefficients, typically the Tanimoto coefficient, based on the numbers of fragments common and not-common to a pair of molecules. Conversely, similarity calculations employing property data have generally involved distance coefficients such as the Euclidean distance, which is widely used in many applications of multivariate statistics. The use of these two types of coefficient is detailed in Figure 2.

A quantitative measure of the degree of diversity in a set of molecules is normally referred to as a *diversity index*, and many such indices have been described in the literature. A common approach is to calculate an index based on the inter-molecular structural similarities for a dataset, as reviewed recently by Waldman et al. [8]. For example, one might take the sum of all the pairwise similarities, or the sum of the similarities for just the nearest neighbour (i.e., most similar) compound for each member of the dataset; in either case, the smaller the sum of the similarities, the greater the degree of diversity. Other quantitative measures of diversity can be obtained from counting the number of different fragments, whether in 2D or 3D, that can be generated from a dataset, and from the number of occupied bins in the partition-based selection schemes discussed in the following section.

A diversity index provides a simple way of comparing the degree of structural heterogeneity in different datasets. For example, a pharmaceutical company looking to increase the range of structural types in its corporate database might compare the diversities of datasets offered by several different external compound suppliers; or a synthetic chemist might decide which of several combinatorial libraries to synthesise based on their calculated diversities. It must, however, be remembered that what is being calculated here is a measure of *structural* diversity, whereas the rationale for carefully selecting the compounds to be synthesised and tested is to ensure *biological* diversity: either as many different activities as possible in the case of broad screening libraries, or as wide a range of activity against a specific biological target, e.g., an IC_{50}, in the case of focused libraries. It is a common working assumption in medicinal chemistry that similar compounds exhibit similar activities, and that changes in structure are reflected in changes in activity, an assumption that is referred to as the *similar property* [9] or *neighbourhood* [10] principle; however, this principle is just a helpful

rule-of-thumb to which there are innumerable exceptions, and thus structural diversity alone cannot be expected to ensure the identification of useful leads.

Selection Of Database Subsets

The similarities or distances obtained as above often provide the principal input to the various methods that are available for selecting a structurally diverse set of compounds [11]. Early approaches to rational compound selection were based on the use of cluster analysis. Cluster-based selection involves applying a clustering method to a set of molecules, yielding clusters that exhibit a high degree of both intra-cluster similarity and inter-cluster dissimilarity; a diverse subset is then obtained by choosing one compound from each of the clusters in turn. However, cluster-based selection is increasingly being complemented, or even replaced, by partition-based selection and dissimilarity-based selection.

Partition-based (or cell-based) selection requires the identification of a small number of characteristics, these typically being molecular properties that would be expected to affect binding at a receptor site. The range of values for each such characteristic is sub-divided into a set of sub-ranges, and the combinatorial product of all possible sub-ranges then defines the set of cells that make up the partition. Each molecule is assigned to the cell that matches the set of characteristics for that molecule, and a subset is then obtained by selecting one (or some small number) of the molecules from each of the resulting cells. This approach is limited to low-dimensionality datasets but is proving increasingly popular: it is exceedingly fast in operation; it facilitates the comparison of different databases; and it enables the identification of those sections of structural space that are under-represented, or even unrepresented, in a database. Cell-based selection is illustrated in Figure 3.

Cluster-based and partition-based approaches identify a diverse subset by first identifying groups of similar molecules, and then picking one molecule from each cluster or cell, respectively. Dissimilarity-based approaches seek to identify a diverse subset directly, typically by iteratively selecting compounds that are as dissimilar as possible to those that have already been selected. The identification of a subset that is maximally dissimilar is computationally infeasible but approximate procedures have been found to work well in practice and several different algorithms have been described. One such procedure is shown in algorithmic form in Figure 4. Dissimilarity-based selection (and also cluster-based selection) can be used with both low-dimensional datasets and high-dimensional datasets, such as fragment bit-strings.

Thus far, we have not taken any account of the natures of the molecules that are being processed. However, there is little point in selecting compounds that are unlikely to yield potential leads, and ***druglikeness*** or ***drugability*** filtering methods are being increasingly used to focus attention on those compounds that have the greatest ***a priori*** probability of exhibiting the properties of previous leads or known drugs. By far the most common filter is the use of a “Rule of Five”-like criterion [12] based on analysis of a set of known drug molecules. Thus, Lipinski’s much-cited paper suggests that poor

absorption or permeation are likely when at least two of the following criteria are satisfied: there are more than 5 H-bond donors; there are more 10 H-bond acceptors; MW>500; logP>5. Molecules meeting two or more of these criteria are considered to be non-druglike and are hence routinely removed from any database prior to further analysis, as are molecules that contain reactive or toxic fragment substructures. If sets of both known drugs and (assumed) non-drugs are available then data mining techniques can be used to develop rules to classify or to rank molecules in decreasing order of drug-likeness, and low-ranked molecules again removed from further consideration. Examples of techniques that have already been used for this purpose include genetic algorithms, binary decision trees and neural networks [13, 14].

Selection methods of the sort discussed thus far can be used in many ways. Perhaps the simplest is to select a diverse subset of an entire database, such as a company's corporate compound collection or a publicly available database such as the *Available Chemicals Directory*. However, this *cherrypicking* approach is not appropriate when combinatorial libraries, rather than individual molecules, are required: in this case, a choice needs to be made between *reactant-based* and *product-based* selection algorithms. These three types of library design procedure are shown in Figure 5.

Design of Combinatorial Libraries Using SELECT

We will conclude this short review by describing a program that we have developed, called SELECT, for product-based selection of drug-like combinatorial libraries that takes direct account of the combinatorial constraint by means of a genetic algorithm (or GA). A GA is a simple, but powerful, tool for generating good approximate solutions to combinatorial optimisation problems, even if they have massively large solution spaces. GAs have proved to be applicable to a broad range of problems in chemoinformatics [15] including that of designing a maximally diverse combinatorial library. Such a task is, in principle, extremely simple: choose one of the diversity indices mentioned above, such as the sum of the pairwise similarities for the molecules in a dataset; systematically generate each possible subset from the pool of available molecules; and then choose that subset with the largest value for the diversity index. The problem with this dissimilarity-based approach is that there are no less than

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

possible subsets of size n molecules that can be chosen from a dataset of size N molecules; this number is totally infeasible for all but the smallest values of n and N.

The GA in SELECT carries out a search of the space of possible subsets, finally returning that subset with the largest value of the chosen diversity index that it has been able to identify. SELECT can carry out both reactant-based and product-based selection, but is normally used to search product space; while this is far more demanding of computational resources, it has been found to result in libraries that are more diverse than those resulting from reactant-based approaches [16]. In addition to maximising the structural diversity, SELECT also ensures the druglike nature of the selected molecules by comparing their physicochemical properties with those of reference sets of drug

molecules. For example, when designing ACE inhibitors, the reference set might be all the molecules from the *World Drug Index* (WDI) database that are coded as belonging to this therapeutic class. Then the profile of property values in a possible library suggested by the GA is compared with the corresponding profile in the reference set of WDI molecules. The overall predicted utility of a library, its 'fitness' in GA terminology, is then a combination of the calculated diversity index of the set of molecules comprising that library and of the difference between the properties of those molecules and the properties of the chosen reference set. The effectiveness of the program is illustrated in Figure 6, where it will be seen that simple reactant-based selection often results in libraries with poor physicochemical property profiles. The product-based selection, conversely, has enabled the construction of libraries with profiles that are much more "WDI-like" and that are thus more likely to contain bioactive compounds

In early work [5], the fitness in the SELECT GA was a weighted sum of the diversity and the property difference; while simple in concept, this was rather akin to adding apples and pears, and required very extensive experimentation to find the best relative weights. We have now adopted an improved methodology that allows the combination of very different types of molecular characteristic; for example, one could design a library that was as structurally diverse as possible, involved the cheapest possible reactants, and gave a set of products that were as similar as possible to previously discovered molecules with the bioactivity of interest. Other desirable characteristics such as synthetic feasibility, solubility and blood-brain barrier permeability can equally well be accommodated in this general framework, given sufficiently rapid and effective methods for the calculation of these characteristics for the molecules in the possible libraries suggested by the GA.

Conclusions

Computer methods for representing molecules and for calculating inter-molecular similarities have been used for many years in chemical database systems. Developments of these methods are now playing an important role in the design of combinatorial libraries. Thus far, the principal focus has been the processing of 2D structure representations, but this is starting to change with increasing use being made of 3D-derived descriptors, such as 3-point and 4-point pharmacophores [17]. Perhaps the most exciting development here is the introduction of methods for the site-directed design of focused libraries, where account is taken of the geometry of the binding site of the biological target [18]. The inclusion of site-specific and physicochemical information in a diversity analysis provides a powerful way of increasing the effectiveness of methodologies for the rational design of combinatorial libraries.

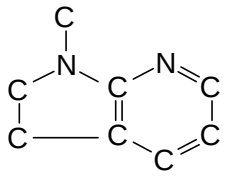
References

1. Willett, P, editor. Special issue on computational methods for the analysis of molecular diversity. Perspect Drug Discov Design 1997;7/8:1-180.
2. Dean PM, Lewis RA, editors. Molecular diversity in drug design. Dordrecht: Kluwer; 1999.
3. Boyd DB, Agrafiotis DK, Martin EJ, editors. Special Issue on combinatorial library design. J Mol Graph Model 2000;18:317-541.

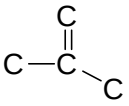
4. Ghose AK, Viswanadhan VN, editors. Combinatorial library design and evaluation: principles, software tools and applications in drug discovery. New York: Marcel Dekker; 2001.
5. Gillet VJ, Willett P, Bradshaw J, Green DVS. J Chem Inf Comput Sci 1999;39:169-177.
6. Dean PM, editor. Molecular similarity in drug design. Glasgow: Chapman and Hall; 1994.
7. Willett P, Barnard JM, Downs JM. J Chem Inf Comput Sci 1998;38:983-996.
8. Waldman M, Li H, Hassan M. J Mol Graph Model 2000;18:412-416.
9. Johnson MA, Maggiora GM, editors. Concepts and applications of molecular similarity. New York: Wiley; 1990.
10. Patterson DE, Cramer RD, Ferguson AM, Clark RD, Weinberger, LE. J Med Chem 1996;39:3049-3059.
11. Willett P. Subset-selection methods for chemical databases. In Ref. 2, pp 115-140.
12. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Adv Drug Deliv Rev 1997;23:3-25.
13. Clark DE, Pickett SD. Drug Discov Today 2000;5(2):49-58.
14. Gedeck P, Willett P. *Curr Opin Chem Biol* 2001;5:389-395.
15. Clark DE, editor. Evolutionary algorithms in computer-aided molecular design. Weinheim: Wiley-VCH; 2000.
16. Gillet VJ, Nicolotti O. Perspect Drug Discov Design 2000;20:265-287.
17. Mason JS, Morize I, Menard PR, Cheney DL, Hulme C, Labaudiniere RF. J Med Chem 1999;42:3251-3264.
18. Siani MA, Skillman AG, Carreras CW, Ashley G, Kuntz ID, Santi DV. J Mol Graph Model 2000;18:497-511.

Peter Willett is Professor of Information Studies in the University of Sheffield, where he heads the chemoinformatics research group. His current research focuses on the use of genetic algorithms, graph theory, and similarity and cluster analysis for the processing of databases of chemical and biological structures . He can be contacted at the Department of Information Studies, University of Sheffield, Western Bank, Sheffield S10 2TN, UK or via email at p.willett@sheffield.ac.uk

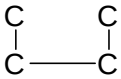
Figure 1. Examples of 2D substructural fragments that can be used to characterise a molecule for the calculation of inter-molecular structural similarity. In these fragments, 'cs' denotes chain-single bond, 'rs' denotes a ring-single bond, 'AA' denotes any atom. In the Ring Fusion fragment, the numbers denote the connectivity, i.e., the number of attached non-hydrogens, for each atom XX in the ring. In the Atom Pair fragment, 'N 0;3' denotes a three-connected nitrogen with no lone pairs, and the '2' denotes the fact that the nitrogen and carbon being characterised are two bonds apart. The reader should note that these are just some of the more common types of fragment used in similarity and diversity analyses, as there are very many different types of fragment definition that could be used.



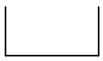
a. Augmented Atom
C cs C rd C rs C



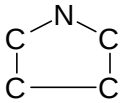
b. Atom Sequence
C rs C rs C rs C



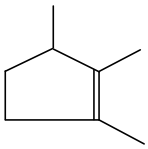
c. Bond Sequence
AA rs AA rs AA rs AA



d. Ring Composition
N rs C rs C rs C rs C rs



e. Ring Fusion
XX3 XX3 XX2 XX2



f. Atom Pair
N 0;3 - 2- C 0;3

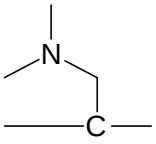


Figure 2. Similarity coefficients used in the calculation of inter-molecular structural similarity: (a) Tanimoto coefficient; (b) Euclidean distance. The Tanimoto coefficient is normally used when molecules are represented by fragment bit-strings. In the calculation, a and b are the numbers of bits that are set in the bit-strings representing the two molecules, I and J, that are being compared, and c is the number of these bits that are set in both bit-strings. The Euclidean distance is normally used when molecules are represented by sets of (normally calculated) physicochemical properties. In the calculation, x_{ik} ($1 \pounds k \pounds n$, where n is the number of different properties being considered) denotes the value of the k-th such property in the molecule I, and similarly for x_{jk} in J.

$$\text{Tanimoto coefficient} = \frac{c}{a + b - c} \qquad \text{(a)}$$

$$\text{Euclidean distance} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2} \qquad \text{(b)}$$

Figure 3. Partition-based selection. Assume that each of the molecules in a database is characterised by its molecular weight (MW) and logP values. Then the molecules can be plotted in a 2D grid, the axes of which represent these two properties. Each cell in the grid defines a specific range of MW and logP values, and all molecules in that cell will have values within these ranges, and can hence be regarded as being chemically equivalent. A diverse subset is then obtained by choosing one molecule from each cell in turn, e.g., that nearest the centre of the cell.

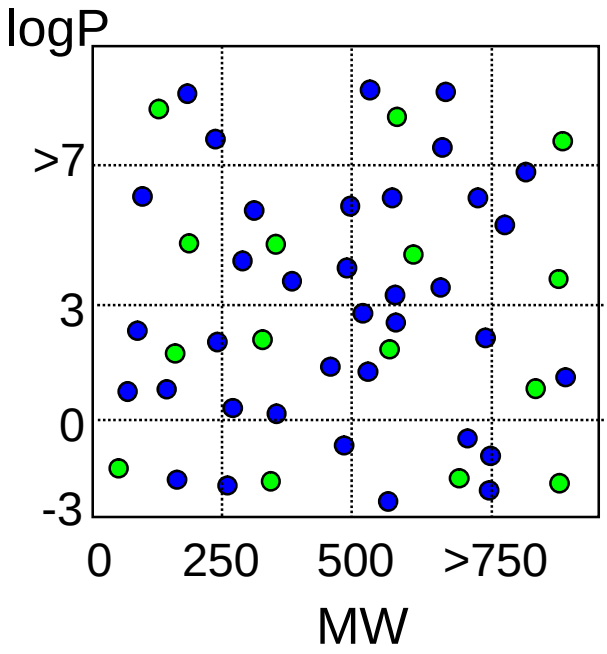


Figure 4. Dissimilarity-based selection. The pseudocode below describes how to select a set of n compounds (called *Subset*) from a larger number of compounds (called *Dataset*). The reader should note that there are very many ways in which a set of compounds can be selected so that they are as dissimilar as possible: thus, one must specify how one chooses the starting compound is chosen in Step 1, how one calculates dissimilarity in Step 2, and how one defines "most dissimilar" in Step 3.

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*.

Figure 5. Compound selection methods. The figure assumes a two-component reaction with pools of N_1 examples of reactant R_1 and N_2 examples of reactant R_2 (e.g., these pools might correspond to carboxylic acids and to primary amines, respectively). Reactant-based approaches to library design involve selecting diverse n_1-member and n_2-members subsets (r_1 and r_2) from R_1 and R_2, respectively, prior to their combination to yield an n_1n_2-member library. Alternatively, the enumeration step in the lower half of the figure involves the *in silico* generation of all possible N_1N_2 products: the selection stage, either by cherrypicking or by invoking the combinatorial constraint, is then applied to this fully enumerated set, which is often referred to as a *virtual library*.

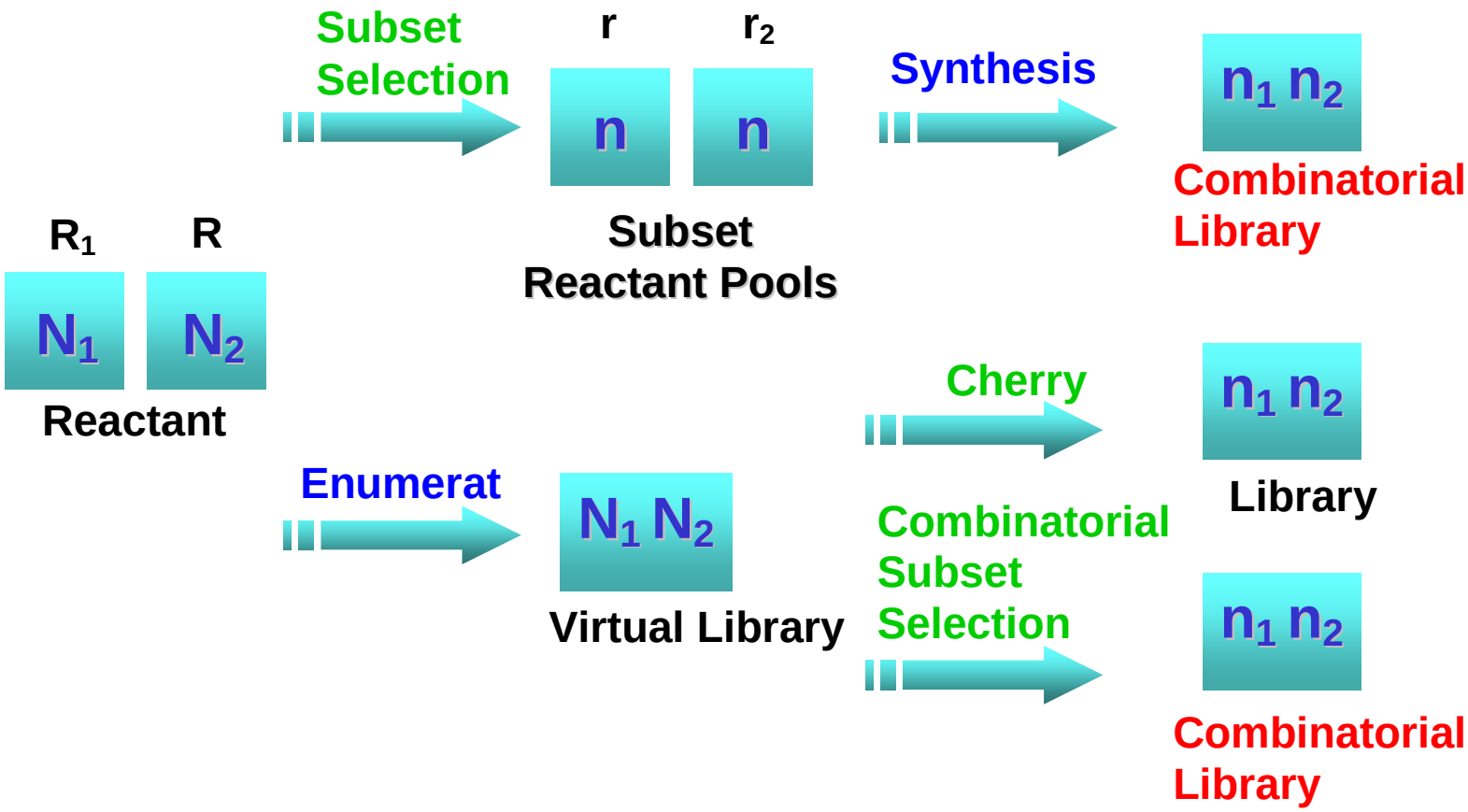
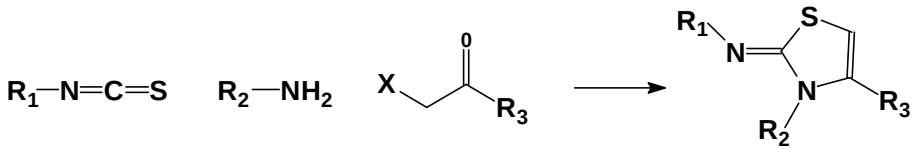
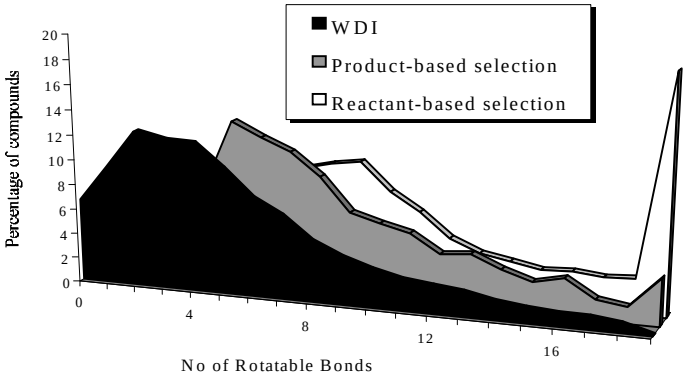


Figure 6. Use of SELECT to design a thiazoline-2-imine library. The combinatorial synthesis is shown in part (a), where the R1, R2 and R3 reactants are isothiocyanates, amines and haloketones, respectively. Sets of 10 isothiocyanates, 40 amines and 25 haloketones were selected at random to give a fully enumerated virtual library of 10000 thiazoline-2-imines, with the molecules represented by Daylight fingerprints and the diversity index being the sum of the pairwise dissimilarities. SELECT was first run to generate diverse sets of reactants (6 isothiocyanates, 10 amines and 15 haloketones) and hence to generate a combinatorial library in reactant space containing 900 thiazoline-2-imines, for which the profile of rotateable bonds was then calculated. SELECT was next run to choose an analogous 900-molecule library in product space, with the library optimised on both diversity and the rotatable bond profile. These SELECT-based profiles were compared with the numbers of rotatable bonds in the molecules in the *World Drugs Index* (WDI) database. The results of these runs are illustrated in part (b).



(a)



(b)