DESIGNING COMBINATORIAL LIBRARIES BY EXPLORING DRUG SPACE

VALERIE J. GILLET

University of Sheffield, Western Bank, Sheffield S10 1EP, UK.
E-mail: V.Gillet@sheffield.ac.uk

Received: 13th July 2000 / Published 11th May 2001

ABSTRACT

The techniques of combinatorial chemistry and high throughput screening are in widespread use in the pharmaceutical and agrochemical industries. During the last few years, many different computational approaches have been developed to select compounds for screening and to design combinatorial libraries. The main approaches are reviewed in the first half of this paper. In the second half, we describe how the library design program SELECT has be used to demonstrate that significant improvements in diversity can be achieved by basing library design in product space rather than in reactant space. A series of experiments are reported involving two combinatorial libraries, three different descriptors and three different diversity indices. Finally, a further significant advantage of performing library design in product space is the ability to optimise multiple properties simultaneously. Thus, SELECT can be used to design libraries that are both diverse and have drug-like physiochemical properties.

INTRODUCTION

Combinatorial chemistry is the process whereby large numbers of compounds are synthesized simultaneously in what are known as combinatorial libraries. The technique, together with the related technology of high-throughput screening, is now used routinely in programs for the discovery of novel bioactive compounds in the pharmaceutical and agrochemical industries. In contrast, traditional approaches to medicinal chemistry involved synthesizing one compound at a time, testing or screening that compound for activity, and then iteratively designing and testing new compounds based on the results. Using traditional methods, a medicinal chemist can synthesize approximately 50 compounds per year. The new technologies, which were introduced in the late eighties and early nineties, have vastly increased throughput so that tens of thousands of compounds can now be made in a single cycle.

Initially the belief was that simply making and testing large numbers of compounds would lead to increased chances of finding actives. However, it soon became apparent that it would not be possible to make all potential compounds due to the combinatorial effect, and nor, in fact, was this desirable. In a typical combinatorial reaction the number of suitable reactants that are available just from commercial sources would result in billions of potential products, which far exceeds the capacity of current combinatorial technologies. For example, Walters et al. [1] report that if all suitable reactants for constructing a benzodiazepine library are extracted from the Available Chemicals Directory the resulting library would consist of in the order of 10^9 compounds. Thus, there is a need to select the compounds that are actually made and tested. The libraries that could potentially be made using all available reactants are referred to as virtual libraries and virtual screening is the process of reducing a

Published in "Chemical Data Analysis in the Large: The Challenge of the Automation Age", Martin G. Hicks (Ed.), Proceedings of the Beilstein-Institut Workshop, May 22nd – 26th, 2000, Bozen, Italy

http://www.beilstein-institut.de/bozen2000/proceedings/gillet/gillet.pdf

virtual library to a practical size for combinatorial synthesis and high-throughput screening. Virtual screening techniques can also be used to select compounds for screening from in-house databases and to determine which compounds should be purchased from external suppliers in compound acquisition programs.

Diversity Analysis and Compound Selection Strategies

Virtual screening, or compound selection, techniques build on the similar property principle, which states that structurally similar molecules are likely to have similar properties. [2] The converse of this suggests that dissimilar molecules will tend to have different properties. This is illustrated in Figure 1, which represents a structure space based on two orthogonal properties. Assuming that the properties are relevant to biological activity, then molecules that are close in the space will tend to have similar biological activity. [3] In terms of a screening experiment, the molecules convey redundant SAR information: they have similar structure and similar activity. In library design and HTS, usually the aim is to screen compounds against a number of different biological targets and a subset of compounds that is evenly spread throughout structure space is likely to maximize coverage of biological activity space. Therefore, there has been a great deal of interest in selecting diverse subsets of compounds that cover as much of the structure space as possible without including redundancy.

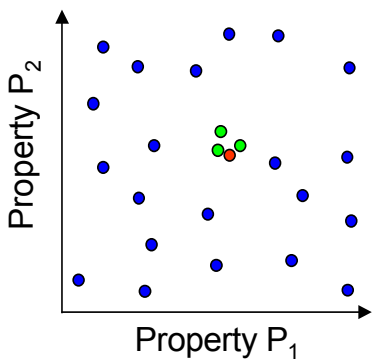

Figure 1: Given an active molecule (red) in a property space that is relevant to biological activity, then according to the similar property principle molecules that are close to it (green) are also likely to

Many diversity measures and subset selection procedures are based on calculations of similarities or dissimilarities between molecules. [4] Consequently, there has been a great deal of interest in measuring structural similarity and dissimilarity [5] and in applying the measures to analyze the diversity of sets compounds [6] and to design diverse combinatorial libraries. Measuring the similarity between two compounds requires that the molecules are described using some numerical descriptors and a coefficient that is used to quantify the degree of similarity between the two sets of descriptors associated with the molecules. [5] The design of diverse libraries requires three major components: the descriptors used to characterize the molecules; a subset selection procedure; and a diversity index that quantifies the degree of diversity in the resulting library.

Descriptors

Many different descriptors have been developed for both similarity and diversity analyses. They have been reviewed extensively (see, for example, [7]) and will be described briefly here. For use in library design, descriptors should have the following characteristics: they should be relevant, that is they should capture structural properties that influence the biological activity of interest; they should be rapid to calculate to enable them to be applied to large datasets; and ideally they should also be chemically interpretable. Descriptors can be categorized as one-, two- or three-dimensional. One-dimensional descriptors are single valued integers or real numbers and they include physicochemical properties, such as molecular

weight, molar refractivity etc., and topological indices, which are indices calculated from the 2D representation of a molecule as a graph. There can be a large number of such descriptors, for example, the Molconn-Z program [8] generates several hundreds of indices, and typically in diversity analyses the number of variables is reduced to a small number that contains most of the information using a technique such as principal components analysis. Two-dimensional fingerprints are probably the most commonly used descriptors in diversity analyses. Three- and four-point pharmacophores are represented as binary vectors or bitstrings, where each bit is set to 0 or 1, depending on whether a particular substructural feature is present or absent from a molecule. Examples include, UNITY [9] and Daylight [10] fingerprints and MACCS screens. [11] Three-dimensional descriptors are also used in diversity analyses, for example, three- and four-point pharmacophores. [12] A pharmacophore point is a substructural feature that is thought likely to influence binding to a receptor, for example, a hydrogen bond donor. They are represented as binary vectors with each bit corresponding to a particular arrangement of pharmacophore points in 3D space. Although they are appealing as descriptors of bioactivity since receptor binding is a 3D event, the calculation of the descriptors is a non-trivial task due to the fact that molecules in general are flexible.

Given a set of molecular descriptors, the similarity (and hence dissimilarity) between two molecules can be calculated using an association coefficient such as the Tanimoto coefficient, which is typically used with binary data such as fingerprints, or a distance measure such as Euclidean distance, which is typically used with physicochemical property data. Association coefficients and distance measures are reviewed by Willett *et al*. [5]

Currently, there is no clear picture as to which descriptors are best. A number of evaluation studies have been performed that tend to suggest that 2D fingerprints are most effective (see, for example, [13] and [14]), however, there is a need for further studies of this kind.

SUBSET SELECTION

Given a set of molecular descriptors and assuming we have some way of measuring diversity *via* a diversity index, then, in theory, the most diverse subset of a given size can be found by generating all possible subsets and calculating the diversity of each one. However, exploring all possible subsets of size n within a dataset of size N requires evaluation of:

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

subsets. This is not computationally feasible for typical values of n and N encountered in library design. Therefore, computationally efficient methods are required to find subsets that are approximate solutions. The subset selection methods that have been applied to selecting diverse subsets can be divided into four different categories: [15] dissimilarity-based compound selection methods; clustering; partitioning and optimization techniques. Each of these methods will be described briefly.

In Dissimilarity-Based Compound Selection (DBCS), [16] heuristics are used to provide good, although non-optimal, solutions to the subset problem. The basic algorithm involves initially selecting the first compound and placing it in the subset and then an iterative loop is entered where in each iteration the compound remaining in the dataset that is most dissimilar to those already in the subset is selected and added to the subset. The algorithm terminates when the required number of compounds has been selected. The various

algorithms developed for DBCS differ in the way the first compound is selected and in the way the dissimilarity between one compound and a group of compounds is measured. For example, the first compound may be selected at random, as the one that is in the center of the dataset, or as the one that is most dissimilar to all the others. The most common ways in which the dissimilarity between one compound and a set of compounds is measured are known as MaxMin and MaxSum. MaxMin selects the compound that has maximum distance to its closest neighbor and MaxSum selects the compound whose average distance to all the compounds in the subset is a maximum. The algorithms are typically used with 2D fingerprints or topological indices as descriptors and the distance coefficients for measuring pairwise dissimilarities are usually the complements of the Tanimoto coefficient and the cosine coefficient.

Clustering [17] is the process of dividing objects into groups, or clusters, so that the objects within a cluster are similar and objects from different clusters are dissimilar. A representative subset can then be chosen by selecting one or more compounds from each cluster. Thus, clustering is an indirect way of selecting a subset since the molecules must first be clustered. The clustering process itself is computationally expensive whereas the subsequent selection process is trivial. As with DBCS, clustering also requires the ability to measure similarities or distances between objects and in subset selection it is most often used with 2D fingerprints. There are many different approaches to clustering but the method that has been found to be most effective is Ward's clustering [13] which is a hierarchical agglomerative method. The computational expense of clustering means that the size of datasets that can be handled is limited.

Partitioning, or cell-based, approaches [12] to subset selection involve firstly defining a low dimensional chemistry space, for example, one that is based on molecular properties such as molecular weight, lipophilicity *etc*. The range of values associated with each property is then divided into a series of bins and the combinatorial product of all bins defines a set of cells that cover the entire space. The molecules are then positioned within the space according to their particular properties. A subset can be chosen by selecting one molecule from each cell. Partitioning methods are sometimes referred to as absolute diversity measures, rather than relative measures since the space is defined independently of the molecules that are positioned within it, unlike clustering, where the clusters are determined by the intermolecular distances themselves. This characteristic of partitioning methods means that it is easy to perform database comparisons, which can be a very useful procedure in library design. Partitioning schemes have been developed for low-dimensional data such as physicochemical properties and also for a new type of descriptor called BCUT descriptors. [12] Partitioning schemes are also used with the vector based, one-dimensional, three- and four-point pharmacophores described earlier. One difference with pharmacophore data compared to physicochemical property data is that a single molecule will typically occupy more than one cell and in some cases an individual molecule can occupy a large number of cells, such molecules are sometimes called promiscuous.

The final category of subset selection algorithms is that of optimization techniques. Several methods have been developed that fall into this category. The methods require definition of a function or diversity index that is to be optimized. Examples of algorithms that have been applied to subset selection include genetic algorithms [18,19,20], simulated annealing, [21] and experimental design techniques. [22] In these algorithms, the function is

calculated many times and hence the complexity of the calculation is restricted. Diversity indices that are used with optimization techniques include distance based indices such as the sum of pairwise dissimilarities [18] and the number of distinct pharmacophores [19] covered by a subset of compounds.

most combinatorial library design efforts to date have been based on reactant-based selection. The methods assume that by maximizing diversity in the reactants, maximum diversity in the product molecules will be achieved. However, recent evidence suggests that significantly more diverse libraries can be achieved if selection is performed in product space [18,23,24,25].

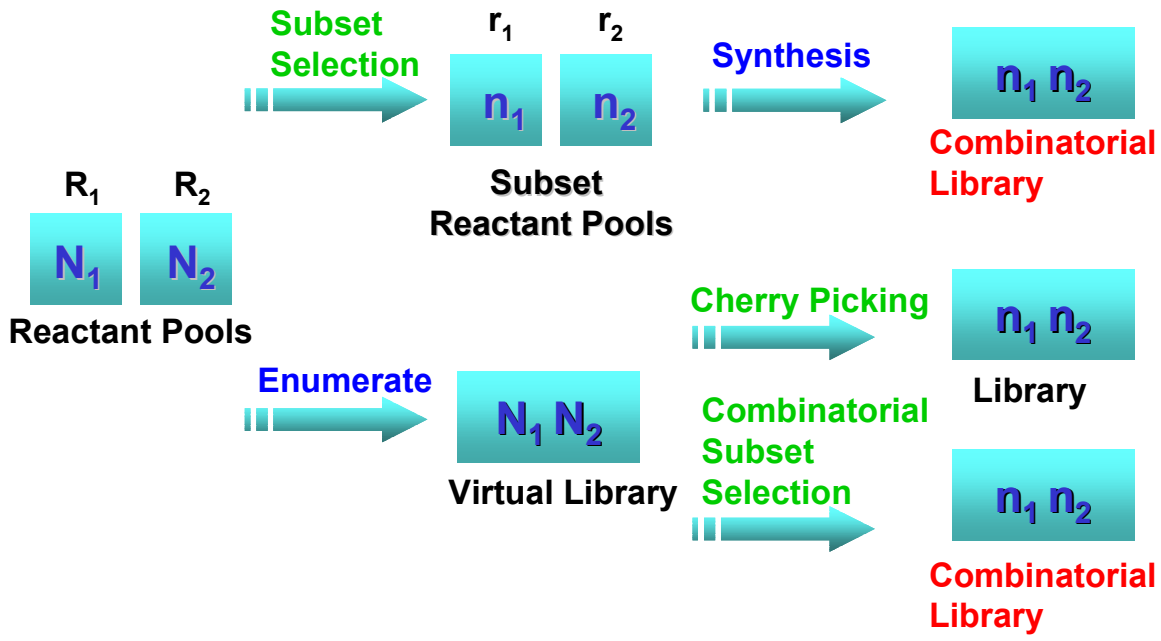

Figure 2: Three different strategies are available for designing diverse combinatorial libraries. They are reactant-based selection, shown in the top half of the figure; cherry-picking in product space; and combinatorial subset selection in product space, known as product-based selection.

COMBINATORIAL LIBRARY DESIGN

The discussion so far has concentrated on strategies for the selection of subsets of compounds with particular reference to the selection of HTS sets from, for example, in-house databases, and the selection of compounds to purchase from external suppliers. In combinatorial library design, any of the techniques already described can be applied directly to choose subsets of reactants from those that are available for use in a combinatorial synthesis. The subsets of reactants are then used to build a combinatorial library in a process known as *reactant-based selection*. This approach is shown schematically in the top-half of Figure 2. Indeed,

In product-based library design a virtual library is enumerated using all available reactants, as shown in the bottom-half of Figure 2. The simplest way of performing product-based selection is to apply any of the techniques described previously in a process known as *cherry-picking*. This approach, however, is synthetically inefficient as far as combinatorial chemistry is concerned since it does not take into account the combinatorial constraint and hence is highly unlikely to result in a combinatorial library. The synthetic inefficiency of cherry-picking is shown in Figure 3, where a two-component combinatorial reaction is represented by a two-dimensional array. The rows represent the reactants

	Y_1	Y_2	Y_3	Y_4	Y_5
X_1	X_1Y_1	X_1Y_2	X_1Y_3	X_1Y_4	X_1Y_5
X_2	X_2Y_1	X_2Y_2	X_2Y_3	X_2Y_4	X_2Y_5
X_3	X_3Y_1	X_3Y_2	X_3Y_3	X_3Y_4	X_3Y_5
X_4	X_4Y_1	X_4Y_2	X_4Y_3	X_4Y_4	X_4Y_5
X_5	X_5Y_1	X_5Y_2	X_5Y_3	X_5Y_4	X_5Y_5

Figure 3: A two component combinatorial library is represented as a two-dimensional array with the reactants in one pool represented by the rows and the reactants in the second pool represented by the columns. A cherry-picked library consisting of four molecules is shown highlighted in red. The 4 × 3 combinatorial library that contains these four molecules is shown in blue.

in one pool, the columns represent the reactants in the second pool and the elements represent the full virtual library. A cherry-picked subset is equivalent to picking compounds from anywhere in the array, for example the subset of four compounds that are highlighted. Synthesizing these four compounds combinatorially would require synthesis of 12 products in a 4 × 3 library.

A combinatorial subset can be selected by intersecting the rows and columns of the matrix, as shown in Figure 4 where the products of a 2 × 2 combinatorial subset library are highlighted. Generating all possible combinatorial subsets in order to find the most diverse is then equivalent to

	Y_1	Y_2	Y_3	Y_4	Y_5
X_1	X_1Y_1	X_1Y_2	X_1Y_3	X_1Y_4	X_1Y_5
X_2	X_2Y_1	X_2Y_2	X_2Y_3	X_2Y_4	X_2Y_5
X_3	X_3Y_1	X_3Y_2	X_3Y_3	X_3Y_4	X_3Y_5
X_4	X_4Y_1	X_4Y_2	X_4Y_3	X_4Y_4	X_4Y_5
X_5	X_5Y_1	X_5Y_2	X_5Y_3	X_5Y_4	X_5Y_5

Figure 4: A combinatorial subset can be selected by intersecting the rows and columns of the matrix.

permuting the rows and columns of the matrix in all possible ways. However, matrix manipulation of this sort represents an enormous search space and, in practice, investigating all possible combinatorial subsets is infeasible for real library design problems. Once again, an approximate solution can be found by using an optimization technique. We have implemented a genetic algorithm (GA) that is able to select a combinatorial subset from a full virtual library of products, within the program called SELECT. [24] In SELECT, each chromosome of the GA encodes a combinatorial subset in the form of lists of the reactants that make up the library. The fitness function of the GA involves enumerating the sub-library and measuring its diversity. The GA iterates through generations using crossover, mutation and roulette wheel selection until it converges on an optimally diverse library. Diversity can be measured using a number of different molecular descriptors and diversity indices, for example, Daylight fingerprints and the sum-of-pairwise dissimilarities using the cosine coefficient.

SELECT has been used to compare the diversity that can be achieved with reactant-based selection relative to product-based selection. [18,23] The libraries that were examined are a two-component amide library (Figure 5) where the virtual library of 10,000 products is built from 100 amides and 100 carboxylic acids, and a three-component thiazoline-2-imine library (Figure 6), also of 10,000 products, which is built from 10 isothiocyanates, 40 amines and 25 haloketones.

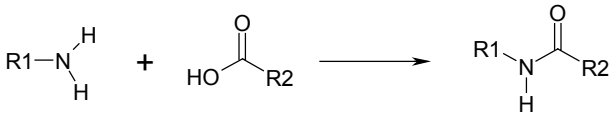

Figure 5: The amide library.

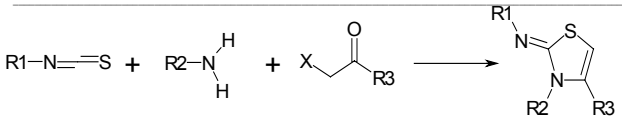

Figure 6: The thiazoline-2-imine library.

Index	Descriptors	Reactants	Products	Min	%Δ
SUM_{COS}	Daylight	0.565 (0.002)	0.586 (0.002)	0.356	9.4
SUM_{TAN}		0.715 (0.002)	0.744 (0.002)	0.522	12.5
NN		0.253 (0.003)	0.305 (0.001)	0.045	20.1
SUM_{COS}	UNITY	0.552 (0.002)	0.566 (0.002)	0.339	5.9
SUM_{TAN}		0.715	0.727	0.507	5.5
NN		0.243	0.294	0.045	20.5
SUM_{COS}	Molconn-Z	0.278 (0.001)	0.288 (0.000)	0.121	6.5
SUM_{TAN}		0.451	0.470	0.217	7.5
NN		0.107	0.150	0.036	37.7

Table 1: Reactant-based versus product-based diversities for 30 × 30 amide libraries selected from a full virtual library of 100 × 100. The column headed *Min* gives the diversity calculated when SELECT was run to find combinatorial subsets with minimum diversity. The final column, %Δ, gives the percentage difference in diversity between product-based and reactant-based selection relative to the range of values possible (calculated by subtracting the *Min* diversity from the Product diversity).

In both cases the reactants were selected at random from the SPRESI database. [26] The experiments were performed for three different types of descriptors, namely 1024 bit Daylight fingerprints, 992 bit UNITY fingerprints, and 538 Molconn-Z parameters that were standardized in the range 0-1. Three different diversity indices were used, namely, the sum-of-pairwise dissimilarities using the cosine coefficient, the sum-of-pairwise dissimilarities using the Tanimoto coefficient, and the average nearest neighbor distance using the Tanimoto coefficient. The results are shown in Tables 1 and 2 for the amide library and the thiazoline-2-imine library, respectively. In all cases it can be seen that product-based designs result in more diverse libraries than do reactant-based designs.

The effect is more pronounced over all the descriptors and metrics for the three-component thiazoline-2-imine library. Unlike reactant-based selection, product-based selection takes into account the relationships between reactants in different pools and hence it is reasonable to expect that the relative effectiveness of product-based selection should increase with the number of reactant pools.

Index	Descriptor	% Δ
SUM_{COS}	Daylight	24.8
SUM_{TAN}		22.3
NN		34.6
SUM_{COS}	UNITY	12.9
SUM_{TAN}		8.0
NN		35.6
SUM_{COS}	Molconn-Z	12.6
SUM_{TAN}		11.4
NN		49.2

Table 2: The percentage difference between reactant-based and product-based diversities is reported for 6 × 10 × 15 thiazoline-2-imine libraries selected from a full virtual library of 10 × 40 × 25.

The effectiveness of product-based selection versus reactant-based selection using SUM_{COS} or SUM_{TAN} as the diversity index is more pronounced for Daylight fingerprints than for UNITY fingerprints or Molconn-Z parameters. Daylight fingerprints are based on calculating the paths of up to 7 atoms within a molecule. When they are calculated for a product molecule there are likely to be several paths that span reactants that originate in different pools, thus there will be parts of the fingerprint that are unique to the product molecule and that are not found in its constituent reactants. This is especially the case for the three-component library. Thus it is not surprising that better results can be achieved by

performing the analysis in product-space. UNITY fingerprints, however, also include some structural keys that record the presence or absence of particular fragments. The structural keys tend to be more localized than the path-based fragments and hence there will be fewer bits that arise in the product molecules only. It is more difficult to explain the performance seen with the Molconn-Z parameters since they encompass a huge range of types of molecular descriptor.

The difference between product-based and reactant-based selection is most marked for the NN diversity index. Combinatorial libraries tend to contain clusters of closely related compounds since each reactant in a reactant pool exists in a product molecule with every reactant in the other pools. The presence of closely related compounds can still result in relatively high diversity values using both the SUM_{COS} and SUM_{TAN} indices. [27] However, the presence of clusters of compounds will result in low diversity values according to the NN index, which prefers an even distribution of compounds. Thus, maximizing the NN index in product space is likely to produce a better spread of compounds throughout the space than can be achieved by just considering the reactants alone.

DRUG-SPACE

Early libraries designed on the basis of diversity did show increased rates of finding hits. However, inspection of the hits revealed that they tended to have undesirable properties as far as potential drug candidates are concerned. For example, they tended to have high molecular weights, to be too lipophilic, or to be insoluble. [28] In fact, it appears that maximizing diversity tends to bias the molecules in libraries away from desired ranges of these properties. Thus, the emphasis in library design has now shifted towards designing libraries that, while still diverse, contain compounds constrained to have drug-like physicochemical properties.

One way in which this is attempted is by applying preliminary filters to eliminate non-drug-like molecules from the reactant pools, for example, removing toxic and reactive groups; compounds

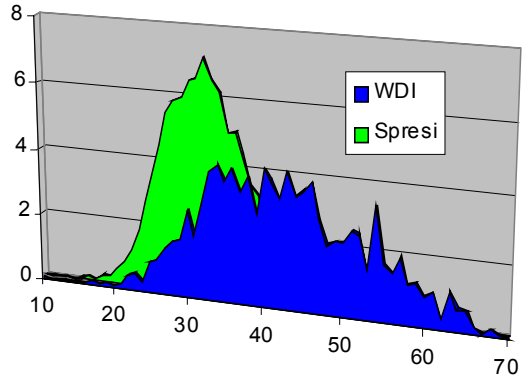

Figure 7: Drug-like weights are used to discriminate between compounds in WDI and compounds in the SPRESI database.

with a large number of rotatable bonds; and compounds with high molecular weights. Some subset selection techniques now use additional properties within the design, for example, tailored D-optimal design [28] and the use of secondary properties to select compounds from cells in a

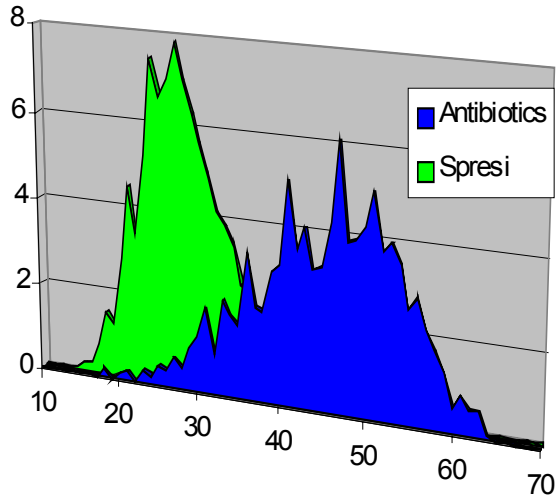

Figure 8: Weights have been derived to discriminate between compounds having antibiotic activity from non-drug-like molecules as found in SPRESI

partitioning procedure.

Recently, several more sophisticated approaches have been described that attempt to predict drug-likeness. [29-32] These methods have their basis in the fact that physicochemical properties are distributed differently in databases of drug-like molecules relative to non-drug-like molecules. [29] An example of this type of approach is the bioactivity profiles approach we have developed, [29] where a series of weights is derived that represent different values of physicochemical properties. The weights can be used to score and rank molecules according to their ability to discriminate between active and inactive compounds. Optimum weights are found using a GA. Figure 7 shows the results of applying drug-like weights to discriminate between molecules in the World Drugs Index, [33] which represents drug-like molecules, from molecules in SPRESI which represents non-drug-like molecules. The method

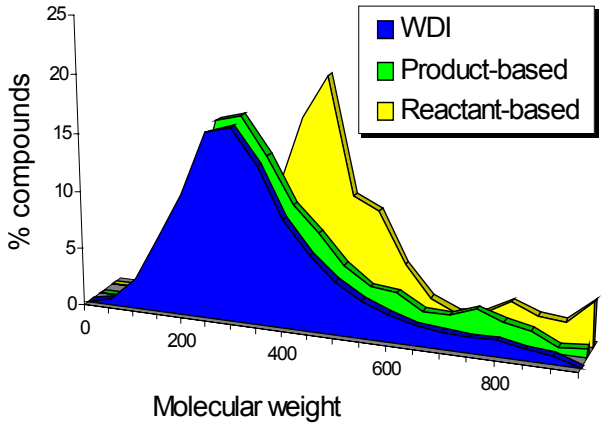

Figure 9: The molecular weight profiles of amide libraries designed using reactant-based selection (in yellow) are compared with libraries that are optimized in product-space (green) and the profile of molecular weights found in WDI (blue).

can be tailored for different classes of activity, for example, in Figure 8, weights have been derived to discriminate antibiotics from non-drugs. This method, and similar methods, can be used to rank datasets for screening so that the compounds that are predicted to be drug-like are screened first, and they can also be used to choose compounds from external suppliers in compound acquisition programs. In the library design context they could be used to choose drug-like reactants, however, they are less suited to product-based design since they do not take account of the combinatorial constraint.

Designing Drug-Like Combinatorial Libraries

We have extended the SELECT program to perform multi-objective optimization in product-space in order that libraries can be designed on multiple properties simultaneously. The fitness function of the GA now consists of a weighted sum as shown:

$$f(n) = w_1.diversity + w_2.complementarity + w_3.cost + w_4.property1 + w_5.property2 \ldots$$

The objectives on library design would typically include diversity along with a number of other properties. The complementarity term can be used to design a library that is complementary to an existing library by maximizing the diversity that would result if the two libraries were merged. The third term represents the cost of synthesizing a library which can be estimated from the cost of the individual reactants that constitute the library. The remaining terms can be used to tailor the physicochemical property profiles of a library. The properties of a library are optimized by comparing the distribution of the property within a library with the distribution of the same property in some reference collection, for example, this could be a collection of drug-like molecules such as those found in the WDI. The weights are user-definable and are usually set to maximize diversity and complementarity while minimizing normalized values of cost and the RMSD between the profile of the properties within the library and the reference profiles.

The effect of multi-component optimization can be seen in Figure 9 where the molecular weight profile of an amide library selected by performing reactant-based selection on diversity alone is show in yellow. The profile of a library selected by performing product-based selection based on diversity and molecular weight simultaneously is shown in green. The molecular weight profile is optimized relative to the profile of molecular weight found in WDI, which is shown in blue. It can be seen that reactant-based selection often results in libraries with poor physicochemical properties. The product-based selection, conversely, has enabled the design of libraries with profiles that are much more WDI-like and that are thus more likely to contain bioactive compounds.

CONCLUSIONS

Many different approaches to designing diverse libraries have been developed, involving a variety of different subset selection techniques and molecular descriptors. We have shown that product-based selection results in libraries that are more diverse than if selection is performed at the reactant-level. Experience has shown that libraries designed on diversity alone have a tendency to contain non-drug-like molecules and it is now apparent that other criteria should also be taken into account. Product-based designs such as that developed in the SELECT program allow for multiple properties to be optimized simultaneously.

ACKNOWLEDGEMENTS

Thanks are due to John Bradshaw, Darren Green, Orazio Nicolotti, Peter Willett and David Wilton for their contributions to SELECT and the bioactivity profiling work, which was funded by GlaxoWellcome Research and Development with software support provided by Daylight Chemical Information Systems and Tripos Inc.

REFERENCES AND NOTES

[1] Walters, W.P.; Stahl, M.; Murcko, M.A. Virtual Screening – An Overview. *Drug Discovery Today*, **1998**, *3*, 160.

[2] Johnson, M.A.; Maggiora, G.M., Eds. Concepts and Applications of Molecular Similarity, John Wiley, New York **1990**.

[3] Patterson, D.E.; Cramer, R.D.; Ferguson, A.M.; Clark, R.D.; Weinberger, L.E. Neighbourhood Behaviour: A Useful Concept for Validation of "Molecular Diversity" Descriptors. *J. Med. Chem*. **1996**, *39*, 3049.

[4] Gillet, V.J. Background Theory of Molecular Diversity. In *Molecular Diversity in Drug Design*; Dean, P.M.; Lewis, R.A., Eds., Kluwer, Dordrecht **1999**.

[5] Willett, P.; Barnard, J.M.; Downs, G.M. Chemical Similarity Searching, *J. Chem. Inf. Comput. Sci.*, **1998**, *38*, 983.

[6] Willett P. Computational Tools for the Analysis of Molecular Diversity. *Perspect. Drug Disc. Design*, 1997;7/8, 1.

[7] Brown, R.D. Descriptors for Diversity Analysis. *Perspect. Drug Discov. Design*., **1997**, *7/8*, 31.

[8] The MOLCONN-Z software is available from eduSoft LC at URL http://www.eslc.vabiotech.com/

[9] UNITY Chemcial Information Software. Tripos Inc., 1699 Hanley Rd., St. Louis, MO 63144.

[10] Daylight Chemical Information Systems, Inc., Mission Viejo, CA, USA.

[11] MACCS II. Molecular Design Ltd., San Leandro, CA.

[12] Mason, J.S.; Pickett, S.D. Partition-Based Selection. *Perspect. Drug Disc. Design*, **1997**, *7/8*, 85.

[13] Brown, R.D.; Martin Y.C. The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand Binding. *J. Chem. Inf. Comput. Sci*., **1997**, *37*, 1.

[14] Matter H. Selecting Optimally Diverse Compounds from Structural Databases: A Validation Study of Two-Dimensional and Three-Dimensional Molecular Descriptors. *J. Med. Chem*. **1997**, *40*, 1219.

[15] Willett P. Subset-Selection Methods for Chemical Databases. In *Molecular Diversity in Drug Design*, Dean, P.M.; Lewis, R.A., Eds., Kluwer, Dordrecht **1999**.

[16] Lajiness, M.S. Dissimilarity-Based Compound Selection Techniques. *Perspect. Drug Disc. Design*, **1997**, *7/8*, 65.

[17] Dunbar, J.B. Cluster-Based Selection. *Perspect. Drug Disc. Design*, **1997**, *7/8*, 51.

[18] Gillet, V. J.; Willett, P.; Bradshaw, J. The

Effectiveness of Reactant Pools for Generating Structurally-Diverse Combinatorial Libraries. *J. Chem. Inf. Comput. Sci*, **1997**, *37*, 731.

[19] Good, A.C.; Lewis, R.A. New Methodology for Profiling Combinatorial Libraries and Screening Sets: Cleaning Up the Design Process with HARPick, *J. Med. Chem*., **1997**, *40*, 3926.

[20] Brown, R.D.; Martin, Y.C. Designing Combinatorial Library Mixtures Using a Genetic Algorithm. *J. Med. Chem*., **1997**, *40*, 2304.

[21] Agrafiotis, D.K. Stochastic Algorithms for Maximizing Molecular Diversity. *J. Chem. Inf. Comput. Sci.*, **1997**, *37*, 841.

[22] Martin, E.J.; Blaney, J.M.; Siani, M.S.; Spellmeyer, D.C.; Wong, A.K.; Moos, W.H. Measuring Diversity - Experimental Design of Combinatorial Libraries for drug Discovery. *J. Med. Chem*., **1995**, *38*, 1431.

[23] Gillet, V.J., Nicolotti, O. New Algorithms for Compound Selection and Library Design, *Perspect. Drug Disc. Design*, in the press.

[24] Gillet, V.J.; Willett, P.; Bradshaw, J.; Green D.V.S. Selecting Combinatorial Libraries to Optimise Diversity and Physical Properties *J. Chem. Inf. Comput. Sci.*, **1999**, *39*, 169.

[25] Jamois, E.A.; Hassan, M.; Waldman, M. Evaluation of Reagent-Based and Product-Based Strategies in the Design of Combinatorial Library Subsets. *J. Chem. Inf. Comput. Sci*., **2000**, *40*, 63.

[26] The SPRESI database is available from Daylight Chemical Information Systems, Inc., Mission Viejo, CA, USA.

[27] Snarey, M;, Terrett, N.K.; Willett, P.; Wilton D.J. Comparison of Algorithms for Dissimilarity-Based Compound Selection. *J. Mol. Graphics Modelling*, 1997, **15**, 372.

[28] Martin, E.J.; Crichlow, R.W. Beyond Mere Diversity: Tailoring Combinatorial Libraries for Drug Discovery. *J. Comb. Chem*. **1999**, *1*, 32.

[29] Gillet, V.J.; Willett, P.; Bradshaw, J. Identification of Biological Activity Profiles Using Substructural Analysis And Genetic Algorithms. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 165.

[30] Ajay; Walters, W.P.; Murcko, M. Can We Learn to Distinguish Between “Drug-like” and “Nondrug-like” Molecules? *J. Med. Chem.*, **1998**, *41*, 3314.

[31] Sadowski, J.; Kubinyi, H. A Scoring Scheme for Discriminating between Drugs and Nondrugs. *J. Med. Chem.*, **1998**, *41*, 3325.

[32] Wagner, M. van Geerstein, V.J. Potential Drugs and Nondrugs: Prediction and Identification of Important Structural Features. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 280.

[33] The World Drugs Index is available from Derwent Information at URL http://www.derwent.com/