Journal of Computer-Aided Molecular Design, **16:** 371–380, 2002.
KLUWER/ESCOM

Reactant- and product-based approaches to the design of combinatorial libraries

Valerie J. Gillet
Department of Information Studies, University of Sheffield, Western Bank, Sheffield S10 2TN, UK

Introduction

The development of high-throughput screening and combinatorial chemistry in the early to mid nineties has revolutionised modern drug discovery programmes [1]. The techniques allow the rapid synthesis and testing of much larger numbers of compounds than was previously possible. However, despite the increased throughput, the potential number of compounds that are available via combinatorial chemistry is far greater than can be handled by the experimental techniques [2]. Thus methods are required for selecting subsets of compounds for the synthesis of combinatorial libraries that will subsequently be screened for activity.

Early approaches to combinatorial library design were based on maximising diversity. The rationale for diversity lies in the similar property principle [3] which states that structurally similar molecules tend to have similar properties. Accordingly, given a definition of chemistry space that is relevant to biological activity, structures that are close in the space are likely to exhibit similar bioactivity and are therefore redundant in terms of a screening experiment. Thus, a library that maximises coverage of the space should have maximum coverage of biological activity with minimum redundancy [4]. Initial results with large diverse libraries have, however, been disappointing with screens either failing to produce expected hit rates or producing hits that have properties that make them undesirable for drug discovery programmes [5]. For example, the libraries tend to contain compounds that are too lipophilic, too flexible or that have molecular weights that are too high to be useful as lead compounds. Thus, it is now recognised that ADME (absorption, distribution, metabolism and toxicity) properties should also be optimised as well as diversity so that the compounds constitute good start points for further optimisation.

Diverse libraries are appropriate for screening against a range of structural targets or when little is known about the therapeutic target of interest. However, when information is available about the biological target it should be incorporated into the library design process in order to increase the chances of finding hits [6]. Thus, focused libraries are typically designed to occupy restricted regions of chemistry space with the boundaries being defined by the available knowledge. For example, when an active compound is known, the library could be designed to contain molecules that are similar to the known active; when several actives are known the library could be designed to contain molecules that are predicted to be active according to a QSAR model; and when the 3D structure of the target is known, a virtual library could be screened against the target to eliminate molecules that cannot fit into the active site.

In focused library design, it is even more desirable to optimise multiple properties since in addition to matching constraints related to the target, other criteria are often required during lead optimisation, for example, bioavailability and cost of goods. Thus, whether designing diverse or focused libraries it is now recognised that library design is a multiobjective optimisation problem.

There are two main techniques used for designing libraries for combinatorial synthesis, namely reactant-based design and product-based design [7]. In reactant-based design, optimised subsets of reactants are chosen without consideration of the product molecules that will result. In product-based design a virtual product library is enumerated from all available reactants and an optimised combinatorial subset is then selected directly from product space. These two approaches are described in more detail in the next section and their relative strengths and weaknesses are discussed in relation to the design of diverse libraries. The relative effectiveness of the approaches is then illustrated by an example library design scenario. Product-based approaches are especially beneficial when it is desirable to optimise multiple library-based properties simultaneously and product-based approaches to multiobjective library design are discussed including a new approach that is based on a multiobjective genetic algorithm. Finally product-based approaches to the design of targeted or focused libraries are discussed.

Combinatorial library design strategies

Consider a combinatorial reaction with two positions of diversity, for example, a reaction involving the coupling of amines and carboxylic acids via an amide bond and assume that there are 1000 reactants available for each of the variables. A combinatorial synthesis involving all of these reactants would give rise to a library of 10^6 product molecules (1000×1000). However, assume that the aim is to synthesise a diverse combinatorial library of 10^4 compounds constructed as a 100×100 subset. In reactant-based design, each reactant pool is considered independently and in the hypothetical example, the library design process would consist of selecting 100 diverse amines from a possible 1000 and then repeating the selection process for the carboxylic acids. In general, there are

$$\frac{N_i!}{n_i!(N_i - n_i)!}$$

different subsets of size n_i contained within a pool of N_i reactants and in a typical library design $n_i << N_i$ so it is not possible to enumerate all subsets and compare them directly. Thus, approximate methods for selecting subsets must be used.

Compound selection techniques have been used for many years to select subsets of compounds for screening from corporate collections and these methods can be applied directly to select reactants for combinatorial library synthesis. Traditional approaches such as clustering and dissimilarity-based compound selection (DBCS) techniques involve calculating similarities or dissimilarities between pairs of compounds in the dataset. Calculating molecular similarities requires firstly, that the compounds are represented by numerical descriptors and secondly, the application of a similarity coefficient such as the Tanimoto coefficient [8]. Many different descriptors have been used in the application of diversity methods, perhaps the most common being the 2D fingerprint [9].

In clustering, compounds are first grouped into clusters of similar compounds and a diverse subset is selected by choosing compounds across all clusters [10]. In DBCS, a diverse subset of compounds is selected directly in an iterative process where each step involves determining the compound that is most dissimilar to those already included in the subset [11,12]. Several different DBCS algorithms have been developed and they vary in the way in which the first compound is selected and the way in which the dissimilarity of one compound to a set of compounds is measured, with examples of the latter being the MaxMin and MaxSum methods [13].

Partitioning or cell-based methods involve defining a low-dimensional space, for example, a space based on a small number of physicochemical properties such as molecular weight, logP etc [14]. The space is then divided into a number of cells and each molecule is assigned to a cell according to its properties. A diverse subset can then be selected by choosing a compound from each cell.

Other approaches to diverse subset selection include optimisation techniques such as experimental design and simulated annealing [15–18].

In reactant-based design descriptors are calculated for a total of

$$\sum_{i=1}^{R} N_i$$

reactants, where there are R positions of diversity with N_i reactants available for each position. Thus in the hypothetical amide example this means that descriptors are calculated for a total of 2000 reactants consisting of the 1000 amines and 1000 carboxylic acids used as starting points for selection.

The alternative approach to library design is to base the design directly on the products contained within the virtual library. Product-based selection is more computationally demanding than reactant-based selection. Firstly, there are a total of

$$\prod_{i=1}^{R} N_i$$

virtual products (10^6 in the amide example), and typically the product-based approaches involve enumerating the virtual library [19] and calculating descriptors for all product molecules, both of which can be time consuming steps.

The compound selection methods already described for reactant-based selection could be applied to select the required number of molecules (10^4 in the hypothetical example) directly from within product space in a process known as cherry picking. However, the resulting library is likely to be synthetically inefficient from the viewpoint of combinatorial chemistry since the combinatorial constraint is not taken into account. Thus, cherry picking results in what is sometimes known as a sparse matrix as illustrated in Figure 1(a) where a two-component library is represented by a 2D matrix with one set of reactants represented by the columns and the other set represented by the rows. Making the nine highlighted compounds via combinatorial chemistry would require the synthesis of a 7 × 6 library that mostly consists of compounds that are not required. A cherry picked subset of 9 products is shown by the bold circles. A combinatorial subset of nine compounds (3 × 3) is shown in Figure 1(b) and is the result of intersecting 3 rows with 3 columns.

Taking the combinatorial constraint into account, there are

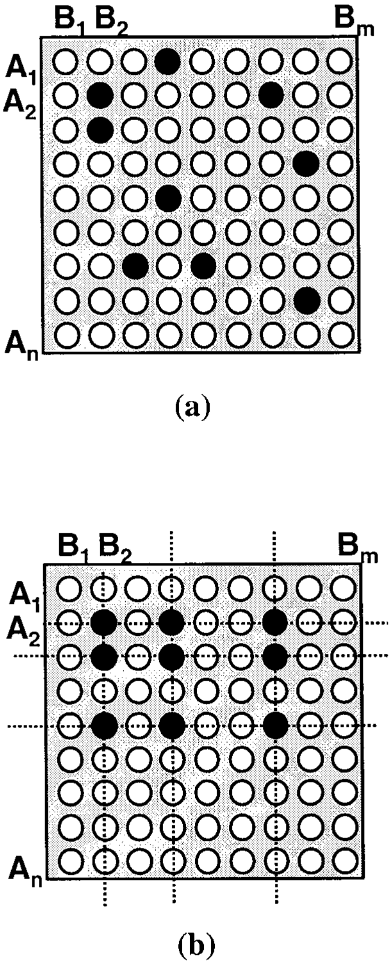
$$\prod_{i-1}^{R} \frac{N_i!}{n_i!(N_i - n_i)!}$$

possible combinatorial subsets to be considered in product-based design as compared with

$$\sum_{i=1}^{R} \frac{N_i!}{n_i!(N_i - n_i)!}$$

for reactant-based design.

The product-based approaches to combinatorial library design are typically implemented using optimisation techniques such as genetic algorithms (GAs) or simulated annealing [7, 20]. For example, the SELECT program [21] is based on a GA where each chromosome of the GA encodes one possible combinatorial subset. Assume a two component combinatorial synthesis in which n_1 of a possible N_1 first reactants are to be reacted with n_2 of a possible N_2 second reactants. The chromosome of the GA contains n_1+n_2 elements, each specifying one possible reactant, and the cross product of the two sets of reactants specifies one of the possible n_1n_2 combinatorial libraries that could be made. The fitness function quantifies the diversity of the sub-library encoded in a chromosome and the GA evolves new potential subset in an attempt to maximise this quantity. Initially, SELECT was designed to optimise distance-based diversity measures such as the sum of pairwise dissimilarities [22], more recent versions include other diversity measures such as a cell-based method.



(a)

(b)

Figure 1.

Other approaches to product-based library design include the GALOPED program developed by Brown and Martin that is also based on a GA [23]. The HARPick program developed by Good and Lewis has both GA and simulated annealing versions and attempts to maximise the number of three-point pharmacophores contained with the library [20]. Mason and Beno describe a simulated annealing approach to optimising diversity both in coverage of BCUT chemistry space and in the number of 4-point pharmacophores present [24]. Other examples of product-based approaches to library design that are based on simulated annealing include the PICCOLO program [25] and the Monte Carlo method developed by Brown et al. [26, 27]

As already indicated, product-based design is much more computationally expensive than reactant-based design. It typically requires the full virtual library to be enumerated and descriptors to be calculated for all potential product molecules. Indeed it may be that in some cases the virtual library is too large to allow full enumeration and thus product-based design is infeasible. (Although the need for full enumeration may not be necessary in the future, for example, Barnard and Downs [28] have recently developed a method for the rapid calculation of descriptors for the products in a virtual combinatorial library that avoids the need for enumeration.) In addition, as already indicated, the subset selection process itself is much more computationally demanding than for reactant-based selection.

Despite the increased computational complexity of performing product-based selection relative to reactant-based selection it has been shown that better optimised libraries result. For example Gillet et al. have shown that more diverse libraries result if selection is performed in product-space rather than in reactant-space [7, 29]. This result was shown to hold for a number of different libraries, descriptors and diversity indices. Similar conclusions were also reached by Jamois et al., although, in some circumstances (that is for some descriptors) they found that reactant-based selection may be comparable to product-based selection [30].

A significant advantage of product-based selection is that it is possible to optimise additional library-based properties such as the physicochemical property profiles of the library as well as the whole molecule properties of the compounds contained within the library. Thus product-based design can be an appropriate strategy when designing libraries over multiple objectives and also in the design of focused libraries as will be described in the following sections. First, the relative effectiveness of product-based library design relative to reactant-based design is illustrated for a diverse 2-aminothiazole library.

Reactant- versus product-based library design

The relative performance of reactant-based selection and product-based selection is illustrated for the two component 2-aminothiazole library shown in Figure 2. Substructure searches were performed on the Available Chemicals Directory [31] to extract suitable α-bromoketones and thioureas, respectively. Each set of reactants was then filtered using the ADEPT software [19] and the following criteria: reactants having molecular weight greater than 300 or more than 8 rotatable bonds were removed; and a series of substructure searches were performed to remove reactants containing undesirable substructural fragments. After filtering there were 74 α-bromoketones and 174 thioureas remaining, which represents a virtual library of 12850 product molecules. Reactant- and product-based approaches were used to design 15 × 30 diverse combinatorial libraries consisting of 15 α-bromoketones and 30 thioureas.

Table 1.

Reactant-based	Product-based	Cherry Picking
0.429	0.448	0.461

Firstly, a library was designed using reactant-based selection. Daylight fingerprints [32] were calculated for the 74 α-bromoketones and 174 thioureas, respectively. Diverse subsets were selected from each reactant pool using dissimilarity-based compound section with MaxSum and the cosine coefficient. The 15 × 30 combinatorial library was enumerated and its diversity was calculated using the sum-of-pairwise dissimilarities, the cosine coefficient and the Daylight fingerprints of the product molecules.

Next, a library was designed in product-space. The full virtual library of 12580 molecules was enumerated and Daylight fingerprints were calculated for all products. The SELECT program was used to find a 15 × 30 combinatorial subset directly, optimised on diversity calculated as the sum-of-pairwise dissimilarities.

The relative diversities achieved using the reactant-based and product-based approaches are shown in Table 1 where it can be seen that, despite the higher computational cost, product-based selection results in greater diversity than does reactant-based selection.

For comparison, an upper bound on diversity was calculated by cherry picking 450 diverse products from the full virtual library using dissimilarity-based compound selection. The increase in diversity achievable when cherry picking is offset by its synthetic inefficiency. For example, in this instance, the 450 diverse products are built from 46 distinct α-bromoketones and 72 distinct thioureas. Synthesising these 450 diverse products via combinatorial chemistry would result in 3312 different products (i.e. a 46 × 72 library).

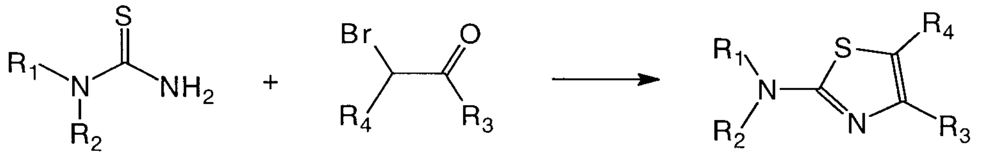


Figure 2.

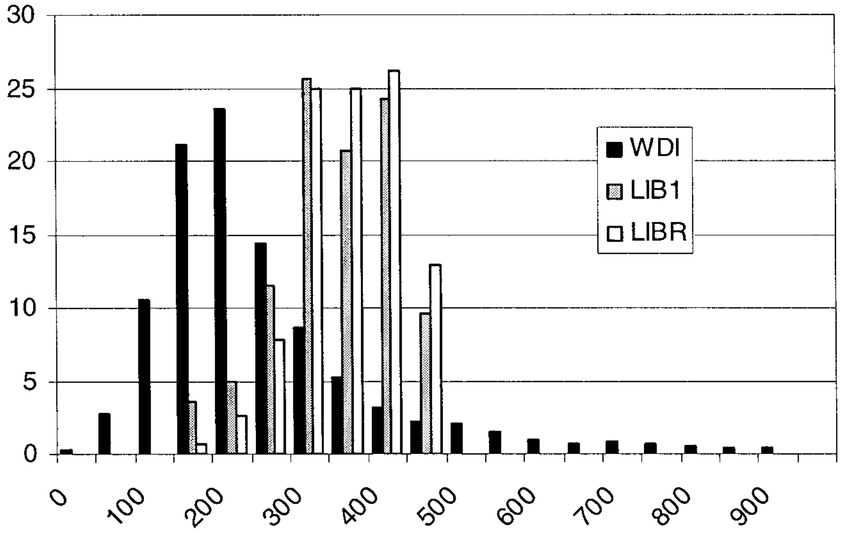


Figure 3.

In Table 1, it can be seen that the product-based method is intermediate in diversity between reactant-based selection and cherry picking with the significant advantage of having maximum synthetic efficiency due to the combinatorial constraint being satisfied.

Similar results with SELECT have been reported previously for a range of different libraries, descriptors and diversity measures [7, 29].

Multiobjective library design

The libraries described thus far have been designed on a single objective, namely diversity, however, as described previously, diverse libraries have a tendency to contain molecules with properties that make them unlikely lead candidates. Figure 3 shows the molecular weight profiles of the product-based library (LIB1) and the reactant-based library (LIBR) superimposed on the molecular weight profile found in the World Drug Index (WDI) [33]. The distribution of molecular weights of compounds in both of these libraries is shifted to significantly higher values than is seen in WDI.

The product-based selection technique implemented in SELECT uses a GA which is a non-deterministic method and may thus lead to different solutions each time it is run. Hence, five different runs were performed with diversity and molecular weight profile recorded for each run. The results are plotted in Figure 4 as diamonds with the molecular weight profile (ΔMW) represented by the RMSD between the profile of the library and the profile found in WDI, scaled to be in a similar range to diversity by dividing by 10. Diversity is plotted on the y-axis with the normal direction reversed so that the improvement in value of each property is towards the origin. It can be seen that each run converges to the same value of diversity however some small variation is seen in the molecular weight profiles although in all cases the distributions are far from ideal.

As already mentioned, product-based library design methods allow additional library-based properties, such as physicochemical property profiles, to be optimised simultaneously with diversity. Thus, the SELECT program has been designed to enable multiple properties of libraries to be optimised via a weighted-sum fitness function. Additional runs were

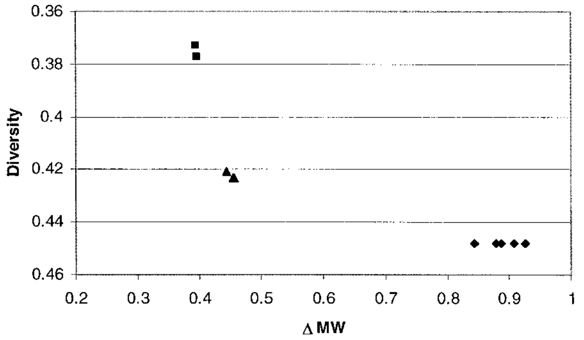


Figure 4.

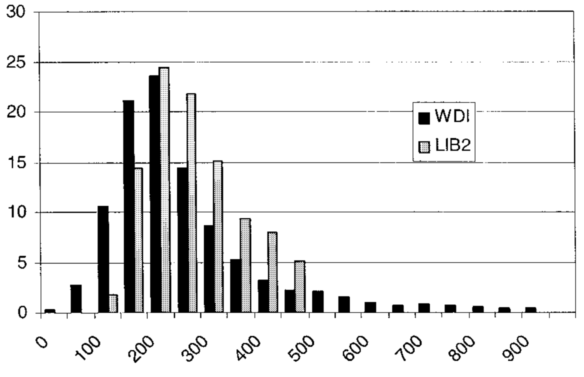


Figure 5.

performed where diversity was optimised simultaneously with molecular weight profile using the fitness function shown:

$$f = w_1(1 - d) + w_2 \Delta MW$$

where d is diversity and minimising $(1 - d)$ is equivalent to maximising diversity, so that the function aims to minimise both terms. The squares in Figure 4 show results obtained when the properties are equally weighted, i.e., $w_1 = w_2 = 1.0$. Here it can be seen that significantly better molecular weight profiles can be achieved at the expense of some diversity. The average diversity found over 5 runs is now 0.397, compared with 0.448 when the library is optimised on diversity alone. Figure 5 shows the molecular weight profile of one of these libraries (LIB2) superimposed on the molecular weight profile found in WDI, where it is clear that the distribution is much more drug-like than that found when diversity alone is optimised using either product or reactant-based selection methods.

Changing the relative weights in SELECT results in a different compromise between diversity and molecular weight as shown by the triangles in Figure 4 for runs with $w_1 = 5.0$; $w_2 = 1.0$. Thus the balance has shifted more in favour of diversity and this may be considered to represent a better compromise between diversity and molecular weight profile.

The weighted-sum approach is easily extended to include additional physicochemical properties, for example, the distribution of ClogP, and any other calculable properties such as cost. A similar approach to multiobjective library design has been implemented in several other library design programs [20, 25–27, 34]. Thus product-based methods allow the design of libraries that, as well as being diverse, are also optimised on additional properties such as the profiles of physicochemical properties. The approach can be used to attempt to ensure that compounds contained within libraries are drug-like as well as diverse. Tailoring combinatorial libraries in this way should increase the chances of finding useful lead compounds.

A new approach to multiobjective library design

Despite the obvious benefits of optimising libraries over a number of properties simultaneously, in practice, finding an appropriate compromise between a number of objectives via a weighted-sum fitness function can be difficult and often requires several trial-and-error runs [35].

Optimisation methods that use a weighted-sum approach to combining multiple objectives, such as the approach used in SELECT, are effectively reducing a multiobjective problem to a single objective. This approach has several limitations. For example, the objectives must be scaled relative to one another and appropriate weights must be assigned to reflect the relative importance of each objective. Both of these can be difficult to achieve, for example, scaling is difficult when the objectives are of different types (such as diversity and cost) and the relative importance of the objectives is often unknown a-priori. The stochastic nature of the search process means that repeating the same experiment using the same weights can produce different results all of which are approximately equal in terms of their overall fitness, however, different values may be obtained for the individual objectives. Varying the relative weights of the objectives, for example, by increasing the importance of diversity will result in a different set of solutions. Figure 4 shows that the objectives diversity and molecular weight profile are in competition with each other for the 2-aminothiazole library with an improvement in molecular weight profile corresponding to a decrease in diversity. A further disadvantage of the weighted-sum approach used in SELECT and other related programs is that each run results in a single solution, when in fact, an entire family of solutions exists all of which can be considered to be equivalent and each of which represents a different compromise or trade-off in the competing objectives.

Multiobjectives Evolutionary Algorithms (MOEAs) are a new class of evolutionary algorithms (EAs) that are designed to handle multiple objectives directly without the need to reduce the objectives to a single weighted-sum objective [36]. MOEAs make use of the population nature of EAs in order to generate an entire family of solutions. The MOGA (MultiObjective Genetic Algorithm) is an example of a MOEA that is based on the GA [37]. In MOGA, each objective is handled independently without the need for summation and hence scaling and the need to apply relative weights. MOGA operates with a population of individuals where each individual represents a potential solution to the problem. The population is refined iteratively in an attempt to identify a Pareto optimal set of solutions. A Pareto-optimal solution, also called a non-dominated solution, is one for which no other solution exists in the population that is better than it in all objectives. The weighted-sum fitness function used in a traditional GA is replaced by fitness based on dominance where an individual is given a rank according to the number of individuals in the population by which it is dominated. The fitness of an individual is then calculated such that all individuals with the same rank have the same fitness and the least dominated individuals are preferred.

The MoSELECT program for multiobjective combinatorial library design was developed to overcome many of the limitations identified in SELECT and is based on a MOGA [38–40]. The chromosome representation is unchanged from SELECT however the ranking and parent selection techniques are now based on dominance calculated over all the objectives as described above. The result of a MoSELECT run is a family of solutions that fall on the Pareto surface all of which are equally valid and each of which represents a different set of compromises between the individual objectives. Full details of the program have been reported previously [38].

Figure 6 shows a family of solutions obtained when the 2-aminothiazole library is optimised simultaneously on diversity and molecular weight profile. The MoSELECT solutions are shown by crosses and are superimposed on the libraries found previously by the SELECT runs, which are represented by triangles. It can be seen that the single run of MoSELECT produces a family of solutions that span the range of different compromise solutions found running SELECT with different relative weights. Thus, MoSELECT allows the relationship between the objectives to be explored in a single run, thereby removing the trial-and-error approach that is required with a weighted-sum. Thus, the library designer can make an informed choice on what represents an appropriate compromise in the objectives.

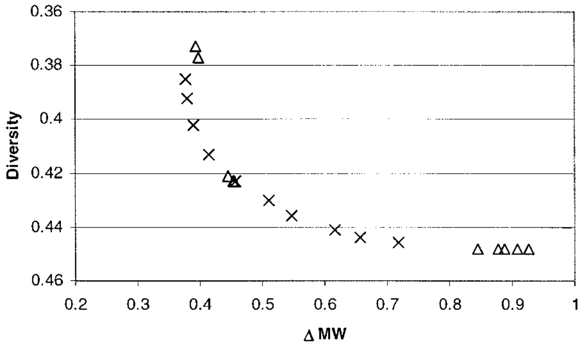


Figure 6.

The MoSELECT approach can be readily extended to include more than two objectives. Additional properties were calculated for the 2-aminothiazole virtual library including number of rotatable bonds and cost (calculated from the price/g quoted in the Available Chemicals Directory). Figure 7 shows the results of designing a library over four objectives, namely, diversity, cost, and profiles of molecular weight and rotatable bonds. The solutions are shown in a parallel co-ordinates representation where a solid line represents a single solution. The solutions are plotted so that the direction of improvement of each objective is towards zero on the y-axis, where zero represents the best value that can be achieved when an objective is optimised independently. The objectives are scaled in order that they can be plotted on the same axis, however, this scaling is not used during the optimisation process itself.

Crossing lines in the parallel co-ordinates view indictate that two objectives are in competition with each other as is clearly the case for diversity and cost. Thus increased diversity corresponds with increased cost. Similarly, as seen for the two-objective case, a drug-like profile of molecular weight is in competition with diversity. Further examples of library design

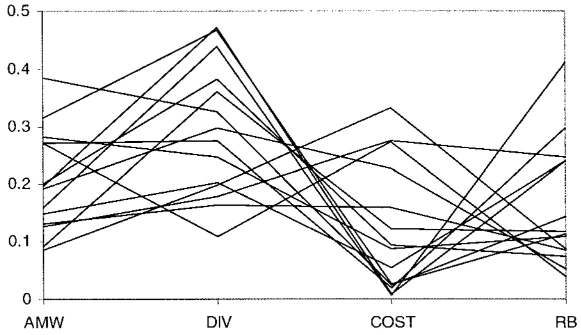


Figure 7.

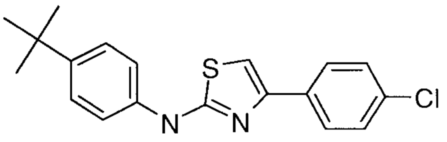


Figure 8.

using MoSELECT have been reported elsewhere [38, 39].

Designing targeted or focused libraries

In targeted or focused libraries it is often whole molecule properties that are of interest, for example, it may be desirable that the libraries contain molecules that are similar to known active compounds. Product-based approaches are hence appropriate for the design of focused libraries. This is illustrated using MoSELECT and the 2-aminothiazole library where the aim is design libraries that contain compounds that are similar to a target compound that was selected from the virtual library at random. The target compound is shown in Figure 8.

A family of libraries of size 15 × 30 were designed optimised on similarity to the target compound simultaneously with cost. Similarity was measured using the sum-of-pairwise similarities using Daylight fingerprints and the Tanimoto coefficient. The solutions found are shown in Figure 9 where similarity is plotted against cost. Again, it is clear that the two objectives are in competition with libraries that contain compounds that are more similar to the target having higher cost. The dashed lines in Figure 9 show the best values that can be achieved for a 15 × 30 combinatorial library when each objective is optimised independently using SELECT. Thus, MoSELECT finds solutions that span the full range of compromise solutions that are possible.

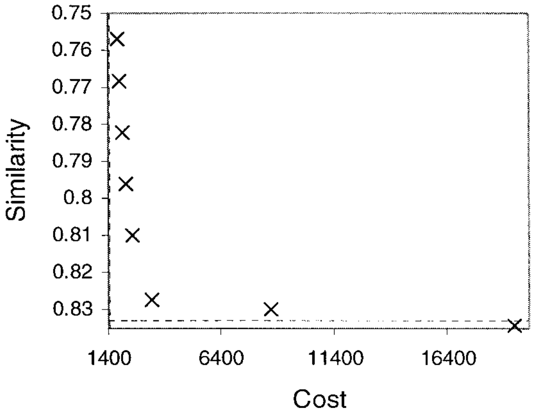


Figure 9.

Other product-based approaches to the design of targeted or focused libraries have been developed that do not take direct account of the combinatorial constraint. These methods have been termed molecule-based methods to distinguish them from library-based methods [41]. For example, Sheridan et al. [41, 42] have developed a method based on a GA where each chromosome of the GA encodes a molecule (rather than a combinatorial library). Individual molecules are optimised according to similarity to a target and once the GA has terminated the entire population of chromosomes is analysed to identify reactants that occur frequently across all the molecules in the population. The frequently occurring reactants can then be used to design a combinatorial library. The algorithm was tested on the construction of tripeptoid libraries where there are 3 positions of variability with 2507 amines available for two of the substitution positions and 3312 for the third position. This represents a virtual library of $\sim$ 20 billion possible tripeptoids. The GA was able to find molecules that were very similar to given target molecules after exploring a very small fraction of the total search space.

The molecule-based method is a relatively fast procedure, especially when optimisation is based on 2D properties, since the fitness function involves a pairwise molecular comparison rather than the analysis of an entire library, as is the case in library-based methods. In these approaches, however, there is no guarantee that building libraries from frequently occurring reactants will result in optimised libraries, although, Sheridan et al. showed that in some cases the molecule-based approaches can be just as effective as library-based approaches. They also showed that basing the optimisation on product molecules is more effective than optimising the reactants.

A similar approach has also been developed in the program Focus-2D [43, 44] where molecules are described using MolconnX topological descriptors and are evolved to be similar to a known target compound or the predicted activity is maximised based on a precomputed QSAR. Both a GA and simulated annealing have been implemented as optimisation techniques. Combinatorial libraries are designed following Monomer Frequency Analysis of the product molecules obtained with monomers, or reactants, that occur frequently in products being considered for the combinatorial synthesis.

An alternative approach to product-based library design has been developed in the PLUMS program [35]. PLUMS seeks a balance between effectiveness and efficiency. The starting point is the combinatorial library that contains all the virtual hits where a hit is a compound that meets some predefined criteria, such as having properties within a given range or that fits a 3D pharmacophore. The method is based on an iterative algorithm where at each iteration the worst monomer is removed successively from the virtual library. The worst monomer is the one that adds least value to the library. Effectiveness is determined by the number of compounds in the library that are virtual hits and efficiency is the ratio of the number of acceptable molecules to the total size of the library. Related approaches have also been described by Stanton et al. [45] and Pickett et al. [46].

Conclusion

Reactant-based and product-based approaches to combinatorial library design have been compared and it has been shown that, despite the higher computational cost, product-based approaches can be more effective at designing libraries that are optimised on diversity. Further advantages of product-based approaches are that additional library-based properties such as physicochemical property profiles can be optimised in multiobjective library design and that whole molecule properties can be otimsed. These advantages have been illustrated by the design of diverse and drug-like libraries and by the design of focused libraries. Recent efforts in computational aspects of combinatorial chemistry have focused on ways of classifying drug-like [47–50] and lead-like compounds [51, 52] and attempts are currently being made to incorporate these methods within combinatorial library design. While diverse libraries are still appropriate when little is known about the biological target, or when libraries are to be screened against a range of targets, the current trend is moving away from very large diverse libraries towards smaller more focused libraries where multiobjective and whole molecule approaches are even more appropriate.

Acknowledgements

The author would like to thank John Bradshaw, Peter Fleming, Darren Green, Illy Khatib, Orazio Nicolotti, and Peter Willett for their contributions to the SELECT and MoSELECT programs described here. The work was funded by GlaxoSmithKline Research and Development with software support being provided by Daylight Chemical Information Systems and Tripos, Inc.

References

1. Leach, A.R. and Hann, M.M., Drug Discovery Today, 5 (2000) 326–336.
2. Walters, W.P., Stahl, M.T. and Murcko, M.A., Drug Discovery Today, 3 (1998) 160–178.
3. Johnson, M.A. and Maggiora G.M. (Eds). Concepts and applications of molecular similarity. Wiley, New York, 1990.
4. Willett P. (Ed.) Perspect. Drug Discov. Design, 7/8 (1997).
5. Martin, E.J. and Critchlow, R.E., J. Comb. Chem., 1 (1999) 32–45.
6. Valler, M.J., Green D., Drug Discovery Today, 5 (2000) 286–293.
7. Gillet, V.J., Willett, P. and Bradshaw, J., J. Chem. Inf. Comput. Sci., 37 731–740.
8. Barnard, J.M., Downs, G.M. and Willett P., J. Chem. Inf. Comput. Sci., 38 (1998) 983–996.
9. Brown, R.D., Perspect. Drug Discov. Design, 7/8 (1997) 31–49.
10. Dunbar Jr., J.B., Drug Discov. Design, 7/8 (1997) 51–63.
11. Lajiness, M.S., Perspect. Drug Discov. Design, 7/8 (1997) 65–84.
12. Gillet, V.J. and Willett, P., In Ghose, A.K. and Viswanadhan, V.N. (eds.) Principles, software tools and applications in drug discovery. Marcel Dekker Inc., New York, 2001, pp 379–398.
13. Holliday, J.D. and Willett, P., J. Biomolec. Screen., 1 (1996) 145–151.
14. Mason, J.S. and Pickett, S.D., Perspect. Drug Discov. Design, 7/8 (1997) 85–114.
15. Martin, E.J., Blaney, J.M., Siani, M.S., Spellmeyer, D.C., Wong, A.K. and Moos W.H., J. Med. Chem., 38 (1995) 1431–1436.
16. Higgs, R.E., Bemis, K.G., Watson, I.A. and Wikel, J.H., J. Chem. Inf. Comput. Sci., 37 (1997) 861–870.
17. Agrafiotis, D.K., J. Chem. Inf. Comput. Sci., 37 (1997) 841–851.

18. Hassan, M., Bielawski, J.P., Hempel, J.C. and Waldman M., Mol. Diversity, 2 (1996) 64–74.
19. Leach, A.R., Bradshaw, J., Green, D.V.S. and Hann, M.M., J. Chem. Inf. Comput. Sci., 39 (1999) 1161–1172.
20. Good, A.C. and Lewis, R.A., J. Med. Chem., 40 (1997) 3926–3936.
21. Gillet, V.J., Willett, P. and Bradshaw, J., J. Chem. Inf. Comput. Sci., 39, (1999) 167–177.
22. Holliday, J.D. Ranade, S.S. and Willett, P., Quant. Struct.-Act. Relat., 14 (1995) 501–506.
23. Brown, R.D., and artin, Y.C., J. Med. Chem., 40 (1997) 2304–2313.
24. Mason, J.S. and Beno, B., J. Mol Graph. Model., 18 (2000) 438–451.
25. Zheng, W., Hung, S.T., Saunders, J.T. and Seibel, G.L., In Atlman, R.B., Dunkar, A.K., Hunter, L., Lauderdale, K. and Klein, T.E. (eds). Pacific Symposium on Biocomputing 2000, World Scientific, Singapore, 2000, pp 588–599.
26. Brown, R.D., Hassan, M. and Waldman, M., In Ghose, A.K. and Viswanadhan, V.N. (eds.) Principles, Software Tools and Applications in Drug Discovery. Marcel Dekker Inc., New York, 2001, pp 301–335.
27. Brown, J.D., Hassan, M. and Waldman, M., J. Mol Graph. Model., 18 (2000) 427–437.
28. Barnard, J.M., Downs, G.M. and von Scholley-Pfab, A., Brown, R., J. Mol. Graphics Model., 18 (2000) 452–463.
29. Gillet, V.J. and Nicolotti, O., Drug Discov. Design, 20, (2000) 265–287.
30. Jamois, E.A., Hassan, M. and Waldman, M., J. Chem. Inf. Comput. Sci., 40, (2000) 63–70.
31. The Available Chemicals Directory is available from MDL Information Systems, Inc., 146000 Catalina Street, San Leandro, CA 94577.
32. Daylight Theory Manual. Daylight Chemical Information Systems, Santa Fe, NM.
33. The World Drug Index is available from Derwent Information, 14 Great Queen St., London W2 5DF, UK.
34. Rassokhin, D.N. and Agrafiotis, D.K., J. Mol Graph. Model. 18 (2000) 427–437.
35. Bravi, G., Green, D.V.S., Hann, M.A. and Leach, A.R., J. Chem. Inf. Comput. Sci. 40 (2000) 1441–1448.
36. Fonseca, C.M. and Fleming, P.J., In De Jong, K. (ed.); Evolutionary Computation; The Massachusetts Institute of Technology, 3, 1995, pp. 1–16.
37. Fonseca, C.M. and Fleming, P.J., In Forrest, S. (ed.) Genetic Algorithms: Proceedings of the Fifth International Conference, Morgan Kaufmann: San Mateo, CA, 1993, pp 416–423.
38. Gillet, V.J., Khatib, W., Willett, P., Fleming, P.J. and Green, D.V.S., J. Chem. Inf. Comput. Sci., In Press.
39. Gillet, V.J., Willett, P., Fleming, P.J. and Green, D.V.S., J. Mol. Graph. Model., In Press.
40. UK Patent Application No. 0029361.
41. Sheridan, R.P., SanFeliciano, S.G. and Kearsley, S.K., J. Mol. Graph. Model., 18 (2000) 320–334.
42. Sheridan, R.P. and Kearsley, S.K., J. Chem. Inf. Comput. Sci., 35 (1995), 310–320.
43. Zheng, W., Cho, S.J. and Tropsha, A., J. Chem. Inf. Comput. Sci., 38 (1998) 251–258.
44. Cho, S.J., Zheng, W. and Tropsha, A., J. Chem. Inf. Comput. Sci., 38 (1998) 259–268.
45. Stanton, R.V., Mount, J. and Miller, J.L., J. Chem. Inf. Comput. Sci., 40 (2000) 701–705.
46. Pickett, S.D., McLay, I.M. and Clark, D.E., J. Chem. Inf. Comput. Sci., 40 (2000) 263–272.
47. Gillet, V.J., Willett, P. and Bradshaw, J., J. Chem. Inf. Comput. Sci., 38 (1998) 165–179.
48. Ajay, Walter, W.P. and Murcko, M.A., J. Med. Chem. 41 (1998) 3314–3324.
49. Sadowski J. and Kuginyi, H., J. Med. Chem., 41, (1998) 3325–3329.
50. Clark, D.E. and Pickett, S.D., Drug Discovery Today, 5 (20000) 49–58.
51. Hann, M.M., Leach, A.R. and Harper, G., J. Chem. Inf. Comput. Sci., 41 (2001) 856–864.
52. Oprea, T.I., Davis, A.M., Teague, S.D. and Leeson, P.D., J. Chem. Inf. Comput. Sci., 41 (2001) 1308–1315.