New Directions in Library Design and Analysis

Val Gillet

Summary

The high costs associated with high throughput screening (HTS) coupled with the limited coverage and bias of current screening collections is such that diversity analysis continues to be an important criterion in lead generation. Whereas early approaches to diversity analysis were based on traditional descriptors such as two dimensional fingerprints a recent emphasis has been on assessing scaffold coverage to ensure that a variety of different chemotypes are represented. Moreover, whether designing diverse or focussed libraries, it is widely recognised that designs should aim to achieve a balance in a number of different properties and multiobjective optimisation provides an effective way of achieving such designs.

Introduction

Although current HTS technologies permit up to one million compounds to be screened in a few weeks, the costs associated with running the screens and the need to replenish samples mean that this is not always desirable [1,2]. Furthermore, it may not always be possible to scale-up the assay to permit HTS. Thus, there is considerable pressure to minimise the number of compounds to be screened. Focussed screening involves the selection of a subset of compounds according to an existing structure-activity relationship, which could be derived from known active compounds or from a protein target site, depending on the available knowledge. Focussed screening is generally desirable; however, it is not always possible, for example, when little is known about the target. In such cases, screening sets are usually designed to be diverse and a sequential screening strategy may be followed. Sequential screening is an iterative process which starts with a small representative set of diverse compounds, the aim being to derive some structure-activity information during the first round of screening which is then be used to select more focussed sets in subsequent rounds of screening. Diversity analysis is also an important criterion when purchasing compounds from external vendors to augment an existing collection: despite a recent growth in corporate screening libraries, the chemical space they cover represents a tiny fraction of the space occupied by drug-like compounds (a typical corporate collection is in the region of 1 to 10 million compounds whereas even conservative estimates of drug-like chemical space are on the order of 10^{13} [3]).

This review focuses on recent developments in library design with particular emphasis on scaffold diversity and combinatorial libraries.

The rationale

Diversity selection has its basis in the similar property principle which states that structurally similar compounds are likely to have similar properties. Thus compounds that are structurally similar to a known biologically active compound are likely to share the

same activity. When considering diverse subsets, the aim is usually to maximise the coverage of structural space while minimising redundancy, i.e., the inclusion of compounds that are so similar that they share the same activity. The similar property principle forms the foundations of rational approaches to medicinal chemistry with numerous examples of its successful application in the literature, however, there are also many counter-examples where a small change in structure can result in the loss of activity, see for example [4].

A common approach to avoiding redundancy has been to use a similarity threshold, for example, a threshold of 0.85 similarity (using 2D descriptors and the Tanimoto coefficient) has been used to reject compounds for purchase [5]. This threshold was based on the observation that if two molecules are 85% similar and one is active, then the other has an 80% chance of also being active [6,7]. However, more recent work has shown a weaker relationship between structure and activity with, on average, only 30% of compounds within 0.85 similarity of an active compounds also being active [8]. This later finding has led to the reconsideration of appropriate thresholds to use when selecting diverse subsets of compounds.

How is diversity measured?

The components necessary to select a diverse subset of compounds include molecular descriptors and a subset selection method. Descriptors include physicochemical properties, topological indices, fingerprint-based descriptors derived from 2D connection tables and 3D conformations. Subset selection methods include dissimilarity-based compound selection which involves calculating pairwise (dis)similarities; clustering which is also based on pairwise similarities; partitioning schemes in which a low dimensional space is defined independent of the compounds themselves which are then mapped onto the space; and optimisation techniques such as simulated annealing and genetic algorithms. Recent comprehensive reviews of diversity analysis are provided by Gorse [9] and Maldonado et al. [10] and a review of descriptors is provided by Glen and Adams [11].

Random vs rational design

The relative merits of random sampling versus more computationally demanding diversity selection has long been debated and the debate continues. For example, Yeap et al report the results of a simulation at Pfizer in which subsets of compounds selected at random from five HTS screens are compared with subsets selected using a cluster-based method [12]. They found that the rationally designed subsets gave higher hit rates than the random subsets. However, contrasting results were found by Schuffenhauer et al. in a recent study on Novartis datasets. Various diversity selection methods, including OptiSim and divisive K-means clustering, were compared with random selections over a wide range of assays and the diversity methods were found to be no better than random at selecting active compounds [13]. It was suggested that this may be due to the limited applicability of the similar property principle (discussed above) coupled with experimental errors in the screening experiment. Schuffenhauer's study also showed that

the OptiSim algorithm has a tendency to bias selections towards those of lower molecular complexity, particularly at low subset sizes, with a corresponding reduction in active molecules. This is thought to be due to the size bias in the Tanimoto coefficient [14] which favours small molecules in dissimilarity-based compound selection [3,14].

Scaffold diversity

A recent emphasis in diversity analysis is on scaffold, or chemotype, diversity. While there is no exact definition of a chemical scaffold, the term is generally used to refer to a common core structure that characterises a group of molecules linked by a common synthetic route [15].

Scaffold hopping has become a popular goal in virtual screening where the aim is to identify active compounds that belong to different lead series from the target compound. Such compounds offer several advantages, for example, they may lead to new patent opportunities and provide alternative lead series should one fail due to poor ADME properties or difficult synthesis. Thus, it has become commonplace for virtual screening methods to be evaluated on their ability to scaffold hop rather than simply on the number of active compounds retrieved. This is often achieved by counting the number of unique molecular frameworks retrieved. The molecular framework was introduced by Bemis and Murko and is defined as the part of a structure that remains after all terminal acyclic atoms have been pruned [16]. A hierarchy of classifications can be devised by progressively discarding atom and bond information from the framework. A related approach has also been developed by Xu and Johnson called Molecular Equivalence Indices, MEQIs [17].

Several reviews of scaffold hopping methods and examples of scaffold hops that were achieved through virtual screening have appeared recently [18-20]. The interest in scaffold hopping has led to many new virtual screening methods which are often considerably more complex than the long established 2D methods involving, for example, 3D descriptors [21-23] and descriptors based on reduced representations of structures [24,25]. However, Sheridan argues that additional complexity may not always be necessary or even desirable, and a better approach may be to use simple descriptors such as atom-pairs and combine information from multiple target structures [26].

Scaffold diversity has also become a popular way of assessing and comparing databases of compounds and provides a different view of the data compared to methods using traditional descriptors such as fingerprints (See Engels et al. for a recent example of database comparison based on 2D fingerprints and the divisive k-means clustering algorithm [27].) Scaffold classification systems can be used to identify under and over represented scaffolds, compare scaffold coverage across different datasets and to analyse HTS datasets.

Several different classification systems have been developed including the molecular frameworks described earlier. A limitation of these, however, is that the presence of a

peripheral ring as a substituent may lead to a compound being classified differently to other compounds in the series. Thus, the HierS classification system was developed to group molecular frameworks hierarchically on the basis of the ring systems contained within them which are obtained by removing the linking bonds [28]. More recently, a unique hierarchical scaffold classification system has been described by Schuffenhauer et al. [29]. Molecular frameworks form the leaf nodes in the hierarchy with higher levels being obtained by iterative removal of rings. Prioritisation rules are used to ensure that peripheral rings are removed first so that, in contrast the HierS classification, a unique classification is obtained. Furthermore, each level of the hierarchy consists of well defined chemical substructures and should therefore be more chemically meaningful than the more abstract representations that are used in the framework approaches.

Scaffold classification systems have been used to compare structural differences between drug-like and nondrug-like compounds [17], screening libraries available from commercial suppliers [30,31], datasets from a variety of different sources (in-house combinatorial library, an HTS dataset, a vendor collection and the World Drug Index) [28] and to analyse natural products [32]. The general aim being to identify over and under represented regions of scaffold space. Such comparisons have highlighted the lack of scaffold diversity that exists in corporate collections [2].

In a rather different approach, Fitzgerald et al. have proposed a method for comparing libraries based on the spatial orientations that are accessible through substitution positions on the library scaffolds [33]. The method is applicable to library scaffolds with three-points of diversity all of which are assumed to contribute to the structure-activity relationship. A conformational search of the scaffold is carried out (with carbon atoms at the substitution positions) and for each conformer, a diversity triangle is generated as the distances between the diversity points. A library is described by the set of diversity triangles generated over all conformations of the scaffold. The method has been applied in a recent survey of combinatorial libraries with three-points of diversity that have been reported in the literature since 1992 [34]. This way of describing a scaffold is similar to the ring system analysis described by Bohl et al. which has been applied to individual molecules with the aim of identifying scaffolds replacements [35].

Scaffold analysis is also being used to mine HTS data with the aim of identifying a more intuitive clustering than can be provided using traditional fingerprints [29,36-39]. Scaffolds that are enriched in active compounds can be used to guide further screening efforts, either through database searching or through combinatorial synthesis and scaffolds that are close to *active scaffolds* in a hierarchical system may provide scaffold hopping opportunities.

Combinatorial library design

Initial efforts in combinatorial library design were aimed at diversity, however, more recently the emphasis has shifted to focussed designs and the design of libraries optimised on multiple properties simultaneously. For example, the importance of ensuring drug-likeness and good ADMET properties as early as possible in a project is

well understood. Moreover, even when designing focussed libraries it is desirable to include an element of diversity to avoid the risk of having multiple hits coming from the same chemical series. Whether designing diverse or focussed libraries, the application of computational filters to remove compounds that have undesirable properties has become widespread and many such filters are available [40,41].

Many approaches have been developed to design combinatorial libraries based on multiple properties. Often they involve the use of optimisation techniques such as genetic algorithms and simulated annealing. The most common way in which multiple objectives are handled is to aggregate them into a single objective by, for example, using a weighted-sum of the individual objectives. See for example [42,43]. A more recent example is described by Le Bailly de Tillegham in which a desirability index is used to combine several properties into a single fitness value [44]. The optimisation method is based on a reagent exchange algorithm which starts with a random set of reagents which are iteratively eliminated and replaced by other reagents.

However, the aggregation approach to combining objectives presents difficulties when the objectives are non-commensurate, for example, diversity and cost, and a trial-and-error approach is usually taken to develop an appropriate weighting scheme. Furthermore, the result of combining the objectives is usually a single solution which represents one particular compromise in the objectives. When the objectives are in conflict, which is usually the case, there can be many different compromise solutions that are all equally valid.

Evolutionary algorithms provide a convenient way to handle multiobjective optimisation since they are population-based and allow a family of different compromise solutions to be explored simultaneously. The objectives are handled independently and the concept of Pareto ranking is used to evolve a family of non-dominated solutions. In Pareto ranking, one population member is said to dominate another if it is better in one objective and at least as good in all others. Many different multiobjective evolutionary algorithms have been developed and they differ in the way in which Pareto ranking is implemented, *inter alia*. For example, in the MOGA the rank of an individual is determined by the number of times it is dominated so that a non-dominated individual is assigned rank 0, an individual dominated by one other population member is assigned rank 1, etc [45]. In the NSGA-II algorithm (Non-dominated Sorting Genetic Algorithm) ranks are assigned in layers [46]. The dominance values of all individuals in the population are calculated and the first non-dominated layer is identified. These individuals are then removed from the population, the dominance values are recalculated and the next non-dominated layer is identified and so on (cf with the layers of an onion).

Multiobjective optimisation has been implemented in the MoSELECT program to design combinatorial libraries over multiple different objectives [47]. For example, it has been applied to the design of focussed libraries where ADME properties have been optimised alongside a similarity criterion [48]. The result is a family of combinatorial libraries, where each library represents a different trade-off in the objectives. The user is then able to choose a library that best suits their needs. It has also been used to explore the trade-off

between library size (number of compounds), configuration (relative numbers of substituents at each position of diversity) and diversity [49]. For example, while it is expected that library coverage should increase with library size, a user is unlikely to know the exact size required for maximum coverage or when the rate of increase in coverage is likely to drop below some threshold. The multiobjective approach allows the full trade-off in size and diversity to be explored, as shown in Figure 1(a) for combinatorial subsets extracted from a 2-aminothiazole virtual library. It has also been shown that improved library designs can be achieved when a library is constructed from more than one combinatorial subset as shown schematically in Figure 1b (Trudi Wright, PhD Thesis. University of Sheffield. 2003). As the number of combinatorial subsets increases the number of products required to achieve maximum diversity decreases considerably (Figure 1c).

Pareto optimisation has also been incorporated into commercially available library design software available from both Accelrys (DS Library Design, Accelrys; URL: http://www.accerlys.com) and Tripos [50]. For example, Soltanshahi et al. describe the application of OptDesign to the design of libraries focussed around known GPCR ligands [51]. The algorithm is based on incremental construction in which the array is built iteratively with a new reagent added each iteration, so that full enumeration is avoided. This approach allows sparse matrices to be designed as well as full combinatorial subsets. A small random sample of reagents is considered in each step and the one that yields the best set of products is chosen for inclusion in the library. The similarity of each potential product to the set of known actives is calculated and the reagents are ranked using Pareto ranking. This is in contrast to MoSELECT where entire combinatorial subsets are evaluated. The best reagent is then chosen based on the Pareto ranks of the products that are generated from it. Thus each reagent is described by a set of points in Pareto space with the number of points taken into consideration determined by the degree of spareness permitted in the final array.

Multiobjective optimisation has also been applied to other applications in chemoinformatics as reviewed recently by Nicolaou et al. [52] For example, Brown et al have developed a method for the de novo design of individual molecules (rather than libraries of molecules) which are optimised on similarity to a set of existing molecules [53]. The approach has also been extended to evolve molecules to fit quantitative-structure property relationship (QSPR) models, such as a solubility model, with the inclusion of indicators of prediction accuracy [54].

Pareto ranking as a tool for data analysis

Pareto ranking is becoming a popular way of analysing data. For example, Figure 2 illustrates a Pareto plot of the property profiles of compounds synthesised and tested in a lead optimisation project (Jeff Loo, MSc Dissertation. University of Sheffield, 2006). Visualisation of the entire set of compounds (shown on the left) provides a useful retrospective evaluation of the project, for example, the relative difficulties associated with optimising the different properties are readily apparent (no compounds met acceptable ranges for properties 4 and 9) and trade-offs in the properties can be identified (between properties 7 and 8 *inter alia*). The profile of the candidate compound is shown in isolation on the right, where although it did not meet the threshold for all the properties, it is Pareto optimal with respect to all other compounds synthesised in the project. This demonstrates the importance of considering multiple properties simultaneously since the sequential application of property filters could have led to the elimination of the candidate, despite it having good values for most of the properties. While a retrospective analysis of this type is instructive, visualisation of an optimisation profile during a project could provide valuable information to help guide decision making. The importance of achieving a balance across a range of criteria is also recognised by other groups [55].

Pareto ranking has also been used to analyse a number of clustering algorithms according to class spread (the average distance of all compounds in a class averaged over all clusters) and number of clusters [15]. A trade-off in these objectives exists with one extreme being all compounds in one cluster (minimum number of clusters; maximum spread) and the other extreme being each compound in its own cluster (maximum number of clusters; minimum spread). Many clustering methods have parameters that are used to control the balance in these two objectives, for example, the Kelley measure [56]. The Pareto analysis showed that the relative performance of different clustering algorithms varied depending on the region of the trade-off surface being considered.

Conclusions

Diversity analysis continues to be a common activity in the design of screening sets, especially when little is known about the target compound. A recent trend in such analyses has been the development of methods to assess scaffold diversity. Such classifications of datasets have emphasised the biases that exists in current screening collections and attention is now turning towards filling the gaps in scaffold space. Increasing scaffold coverage will not necessarily increase hit rates but may result in more series to progress with benefits downstream.

Another area of increasing interest is the diversity available through natural products (NPs). Interest in NPs declined in the 1990s following the introduction of HTS technologies, in part due to high expectations of the new technologies and in part due to the difficulties associated with the isolation and synthesis of NPs. However, the poor performance of HTS has led to a resurgence of interest in NPs. Many comparisons have now been performed of NPs with drug-like compounds, and while NPs do differ from

drugs in several properties, and therefore occupy different region of chemical space, the majority of them do not violate Lipinksi's Rule-of-Five [57]. NPs therefore provide opportunities for exploration of new areas of chemical space relevant to biological activity through NP-derived libraries [32].

1. Schnecke V, Bostrom J: **Computational chemistry-driven decision making in lead generation**. *Drug Discovery Today* 2006, **11**:43-50.
2. Davies JW, Glick M, Jenkins JL: **Streamlining lead discovery by aligning in silico and high-throughput screening**. *Current Opinion in Chemical Biology* 2006, **10**:343-351.
3. Schuffenhauer A, Brown N: **Chemical diversity and biological activity**. *Drug Discovery Today: Technologies* 2006, **3**:387-395.
4. Kubinyi H: **Similarity and dissimilarity: a medicinal chemist's view**. *Perspectives in Drug Discovery and Design* 1998, **9-11**:225-232
5. Martin YC: **What works and what does not: Lessons from experience in a pharmaceutical company**. *Qsar & Combinatorial Science* 2006, **25**:1192-1200.
6. Brown RD, Martin YC: **Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection**. *Journal of Chemical Information and Computer Sciences* 1996, **36**:572-584.
7. Patterson DE, Cramer RD, Ferguson AM, Clark RD, Weinberger LE: **Neighbourhood behaviour: a useful concept for validation of "molecular diversity" descriptors**. *Journal of Medicinal Chemistry* 1996, **39**:3049-3059.
8. Martin YC, Kofron JL, Traphagen LM: **Do structurally similar molecules have similar biological activities?** *Journal of Medicinal Chemistry* 2002, **45**:4350-4358.
9. Gorse AD: **Diversity in medicinal chemistry space**. *Current Topics in Medicinal Chemistry* 2006, **6**:3-18.
10. Maldonado AG, Doucet JP, Petitjean M, Fan BT: **Molecular similarity and diversity in chemoinformatics: From theory to applications**. *Molecular Diversity* 2006, **10**:39-79.
11. Glen RC, Adams SE: **Similarity metrics and descriptor spaces - Which combinations to choose?** *QSAR & Combinatorial Science* 2006, **25**:1133-1142.
12. Yeap SK, Walley RJ, Snarey M, van Hoorn WP, Mason JS: **Designing compound subsets: Comparison of random and rational approaches using statistical simulation**. *Journal of Chemical Information and Modeling* 2007, **47**:2149-2158.
13. Schuffenhauer A, Brown N, Selzer P, Ertl P, Jacoby E: **Relationships between molecular complexity, biological activity, and structural diversity**. *Journal of Chemical Information and Modeling* 2006, **46**:525-535.
14. Holliday JD, Salim N, Whittle M, Willett P: **Analysis and display of the size dependence of chemical similarity coefficients**. *Journal of Chemical Information and Computer Sciences* 2003, **43**:819-828.
15. Schuffenhauer A, Brown N, Ertl P, Jenkins JL, Selzer P, Hamon J: **Clustering and rule-based classifications of chemical structures evaluated in the biological activity space**. *Journal of Chemical Information and Modeling* 2007, **47**:325-336.

16. Bemis GW, Murcko MA: **The properties of known drugs.1. Molecular frameworks**. *Journal of Medicinal Chemistry* 1996, **39**:2887-2893.
17. Xu YJ, Johnson M: **Using molecular equivalence numbers to visually explore structural features that distinguish chemical libraries**. *Journal of Chemical Information and Computer Sciences* 2002, **42**:912-926.
18. Böhm H-J, Flohr A, Stahl M: **Scaffold Hopping**. *Drug Discovery Today: Technologies* 2004, **1**:217-224.
19. Schneider G, Schneider P, Renner S: **Scaffold-hopping: How far can you jump?** *QSAR & Combinatorial Science* 2006, **25**:1162-1171.
20. Zhao HY: **Scaffold selection and scaffold hopping in lead generation: a medicinal chemistry perspective**. *Drug Discovery Today* 2007, **12**:149-155.
21. Rush TS, Grant JA, Mosyak L, Nicholls A: **A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction**. *Journal of Medicinal Chemistry* 2005, **48**:1489-1495.
22. Ahlstrom MM, Ridderstrom M, Luthman K, Zamora I: **Virtual screening and scaffold hopping based on GRID molecular interaction fields**. *Journal of Chemical Information and Modeling* 2005, **45**:1313-1323.
23. Sperandio O, Andrieu O, Miteva MA, Vo MQ, Souaille M, Delfaud F, Villoutreix BO: **MED-SuMoLig: A new ligand-based screening tool for efficient scaffold hopping**. *Journal of Chemical Information and Modeling* 2007, **47**:1097-1110.
24. Wolohan PRN, Akella LB, Dorfman RJ, Nell PG, Mundt SM, Clark RD: **Structural unit analysis identifies lead series and facilitates scaffold hopping in combinatorial chemistry**. *Journal of Chemical Information and Modeling* 2006, **46**:1188-1193.
25. Barker EJ, Cosgrove DA, Gardiner EJ, Gillet VJ, Kitts P, Willett P: **Scaffold-hopping using clique detection applied to reduced graphs**. *Journal of Chemical Information and Modeling* 2006, **46**:503-511.
26. Sheridan RP: **Chemical similarity searches: when is complexity justified?** *Expert Opinion on Drug Discovery* 2007, **2**:423-430.
27. Engels MFM, Gibbs AC, Jaeger EP, Verbinnen D, Lobanov VS, Agrafiotis DK: **A cluster-based strategy for assessing the overlap between large chemical libraries and its application to a recent acquisition**. *Journal of Chemical Information and Modeling* 2006, **46**:2651-2660.
28. Wilkens SJ, Janes J, Su AI: **HierS: Hierarchical scaffold clustering using topological chemical graphs**. *Journal of Medicinal Chemistry* 2005, **48**:3182-3193.
29. Schuffenhauer A, Ertl P, Roggo S, Wetzel S, Koch MA, Waldmann H: **The scaffold tree - Visualization of the scaffold universe by hierarchical scaffold classification**. *Journal of Chemical Information and Modeling* 2007, **47**:47-58.
30. Verheij HJ: **Leadlikeness and structural diversity of synthetic screening libraries**. *Molecular Diversity* 2006, **10**:377-388.
31. Krier M, Bret G, Rognan D: **Assessing the scaffold diversity of screening libraries**. *Journal of Chemical Information and Modeling* 2006, **46**:512-524.
32. Koch MA, Schuffenhauer A, Scheck M, Wetzel S, Casaulta M, Odermatt A, Ertl P, Waldmann H: **Charting biologically relevant chemical space: A structural**

classification of natural products (SCONP). *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**:17272-17277.
33. Fitzgerald SH, Sabat M, Geysen HM: **Diversity space and its application to library selection and design**. *Journal of Chemical Information and Modeling* 2006, **46**:1588-1597.
34. Fitzgerald SH, Sabat M, Geysen HM: **Survey of the diversity space coverage of reported combinatorial libraries**. *Journal of Combinatorial Chemistry* 2007, **9**:724-734.
35. Bohl M, Loeprecht B, Wendt B, Heritage T, Richmond NJ, Willett P: **Unsupervised 3D ring template searching as an ideas generator for scaffold hopping: Use of the LAMDA, RigFit, and field-based similarity search (FBSS) methods**. *Journal of Chemical Information and Modeling* 2006, **46**:1882-1890.
36. Medina-Franco JL, Petit J, Maggiora GM: **Hierarchical strategy for identifying active chemotype classes in compound databases**. *Chemical Biology & Drug Design* 2006, **67**:395-408.
37. Yan SF, King FJ, He Y, Caldwell JS, Zhou YY: **Learning from the data: Mining of large high-throughput screening databases**. *Journal of Chemical Information and Modeling* 2006, **46**:2381-2395.
38. Harper G, Pickett SD: **Methods for mining HTS data**. *Drug Discovery Today* 2006, **11**:694-699.
39. Harper G, Bravi GS, Pickett SD, Hussain J, Green DVS: **The reduced graph descriptor in virtual screening and data- driven clustering of high-throughput screening data**. *Journal of Chemical Information and Computer Sciences* 2004, **44**:2145-2156.
40. Truchon JF, Bayly CI: **GLARE: A new approach for filtering large reagent lists in combinatorial library design using product properties**. *Journal of Chemical Information and Modeling* 2006, **46**:1536-1548.
41. Ghose AK, Herbertz T, Salvino JM, Mallamo JP: **Knowledge-based chemoinformatic approaches to drug discovery**. *Drug Discovery Today* 2006, **11**:1107-1114.
42. Agrafiotis DK: **Multiobjective optimization of combinatorial libraries**. *Journal of Computer-Aided Molecular Design* 2002, **16**:335-356.
43. Zheng WF, Cho SJ, Waller CL, Tropsha A: **Rational combinatorial library design. 3. Simulated annealing guided evaluation (SAGE) of molecular diversity: A novel computational tool for universal library design and database mining**. *Journal of Chemical Information and Computer Sciences* 1999, **39**:738-746.
44. Le Bailly de Tilleghem C, Beck B, Boulanger B, Govaerts B: **A fast exchange algorithm for designing focused libraries in lead optimization**. *Journal of Chemical Information and Modeling* 2005, **45**:758-767.
45. Fonseca CM, Fleming PJ: **Multiobjective optimization and multiple constraint handling with evolutionary algorithms - Part I: A unified formulation**. *IEEE Transactions on Systems Man and Cybernetics. Part A. Systems and Humans* 1998, **28**:26-37.
46. Deb K, Pratap A, Agarwal S, Meyarivan T: **A fast and elitist multiobjective genetic algorithm: NSGA-II**. *IEEE Transactions on Evolutionary Computation* 2002, **6**:182-197.

47. Gillet VJ, Khatib W, Willett P, Fleming PJ, Green DVS: **Combinatorial library design using a multiobjective genetic algorithm**. *Journal of Chemical Information and Computer Sciences* 2002, **42**:375-385.
48. Gillet VJ, Willett P, Fleming PJ, Green DVS: **Designing focused libraries using MoSELECT**. *Journal of Molecular Graphics and Modelling* 2002, **20**:491-498.
49. Wright T, Gillet VJ, Green DVS, Pickett SD: **Optimizing the size and configuration of combinatorial libraries**. *Journal of Chemical Information and Computer Sciences* 2003, **43**:381-390.
50. Clark RD, Kar J, Akella L, Soltanshahi F: **OptDesign: Extending optimizable k-dissimilarity selection to combinatorial library design**. *Journal of Chemical Information and Computer Sciences* 2003, **43**:829-836.
51. Soltanshahi F, Mansley TE, Choi S, Clark RD: **Balancing focused combinatorial libraries based on multiple GPCR ligands**. *Journal of Computer-Aided Molecular Design* 2006, **20**:529-538.
52. Nicolaou CA, Brown N, Pattichis CS: **Molecular optimization using computational multi-objective methods**. *Current Opinion in Drug Discovery & Development* 2007, **10**:316-324.
53. Brown N, McKay B, Gasteiger J: **The de novo design of median molecules within a property range of interest**. *Journal of Computer-Aided Molecular Design* 2004, **18**:761-771.
54. Brown N, McKay B, Gasteiger J: **A novel workflow for the inverse QSPR problem using multiobjective optimization**. *Journal of Computer-Aided Molecular Design* 2006, **20**:333-341.
55. Segall MD, Beresford AP, Gola JMR, Hawksley D, Tarbit MH: **Focus on success: using a probabilistic approach to achieve an optimal balance of compound properties in drug discovery**. *Expert Opinion on Drug Metabolism & Toxicology* 2006, **2**:325-337.
56. Kelley LA, Gardner SP, Sutcliffe MJ: **An Automated Approach for Clustering an Ensemble of NMR-Derived Protein Structures into Conformationally-Related Subfamilies**. *Protein Engineering* 1996, **9**:1063-1065.
57. Wetzel S, Schuffenhauer A, Roggo S, Ertl P, Waldmann H: **Cheminformatic analysis of natural products and their chemical space**. *Chimia* 2007, **61**:355-360.