Computational Techniques for Diversity Analysis and Compound Classification

D.B. Kitchen[1], F.L. Stahura[1, 2] and J. Bajorath*[1, 2, 3]

[1]*Department of Computer-Aided Drug Discovery, Albany Molecular Research, Inc. (AMRI), 21 Corporate Circle, Albany, NY 12212-5098, USA*

[2]*AMRI Bothell Research Center (AMRI-BRC), 18804 North Creek Pkwy, Bothell, WA 98011, USA*

[3]*Department of Biological Structure, University of Washington, Seattle, WA 98195, USA*

Abstract: Molecular similarity and diversity analysis has played a significant role in computer-aided drug discovery for more than a decade. Compound classification methods have also become increasingly important for the design and organization of compound databases and *in silico* screening. Here we review these related methodologies and discuss selected applications.

Keywords: Molecular similarity, Molecular diversity, Chemical descriptors, Compound libraries, Bioactive molecules, Drug-likeness, Clustering, Partitioning.

INTRODUCTION

Drug discovery research over the past 10-15 years has resulted in an explosion of primary biological, chemical, and pharmacological data, which has catalyzed rapid development of informatics approaches in the life sciences. In chemistry, combinatorial and high-throughput synthesis techniques have dramatically increased the size of available compound pools. Large pharmaceutical companies often screen millions of molecules and generate enormous amounts of data, which in turn creates substantial chemoinformatics issues. Computational approaches are not only needed for the design, selection, and organization of new compounds to be made and tested but also for data processing and management and, ultimately, large-scale analysis of structure/property-activity relationships.

Computational approaches to the analysis of molecular similarity and diversity have been at the forefront of chemoinformatics research for more than a decade. The concept of molecular similarity became a focal point of the computational drug discovery community in 1990 [1] and approximately at the same time, algorithms for selection of dissimilar compounds were introduced [2]. Within the next five years, the field of diversity analysis and design evolved [3], spurred on by the need to design chemically diverse combinatorial libraries. In addition, compound classification techniques, in particular, cluster analysis [4], have been intensely studied and applied in computational medicinal chemistry at least since the mid 1980s. For essentially all of these approaches, the ability to computationally compare molecules on a large scale, analyze, and –if possible- quantify their degree of relatedness is of critical importance. Algorithms are needed to explore similarity relationships, the detection of which goes well beyond what a chemist's eye would be able to see or what chemical intuition would tell us, as illustrated in (Fig. **1**). Attractive applications of methods for evaluating molecular similarity or diversity range from the design of diverse compound libraries to the selection of representative subsets. Compound classification methods permit the generation of focused libraries or the identification of novel hits or leads from large compound collections. Such investigations can make important contributions to experimental programs. For example, methods to estimate diversity make it often possible to ensure that synthetic libraries explore new areas of chemistry, while the application of compound classification techniques can greatly reduce the number of compounds selected for biological evaluation and substantially help to analyze the results.

There is little doubt that diversity and classification methods will continue to play an important role in chemoinformatics research and drug discovery. Since these approaches are conceptually related to each other, we review them here in context and discuss some representative applications.

DESCRIPTORS

Assessing molecular similarity and diversity or establishing relationships between compounds (the basis for classification) generally requires the definition of chemical references spaces. A key question is how to represent a collection of compounds on a computer to best reflect their similarity or diversity and group them accordingly. This question can not be unambiguously answered because the generation of chemical spaces for molecular comparisons is always subject to definition and underlying assumptions. Different chemical reference spaces typically produce different compound distributions. However chemical space is defined, its generation for computational analysis stringently depends on the selection of descriptors of molecular structure and properties. Therefore, we begin our discussion with an overview of different chemical descriptors that are being used for these and other purposes. Table **1** summarizes different types of descriptors, as discussed in the following.

*Address correspondence to this author at the AMRI Bothell Research Center (AMRI-BRC), 18804 North Creek Pkwy, Bothell, WA 98011, USA; Tel: (425) 424-7297; Fax: (425) 424-7299; E-mail: jurgen.bajorath@albmolecular.com

1389-5575/04 $45.00+.00

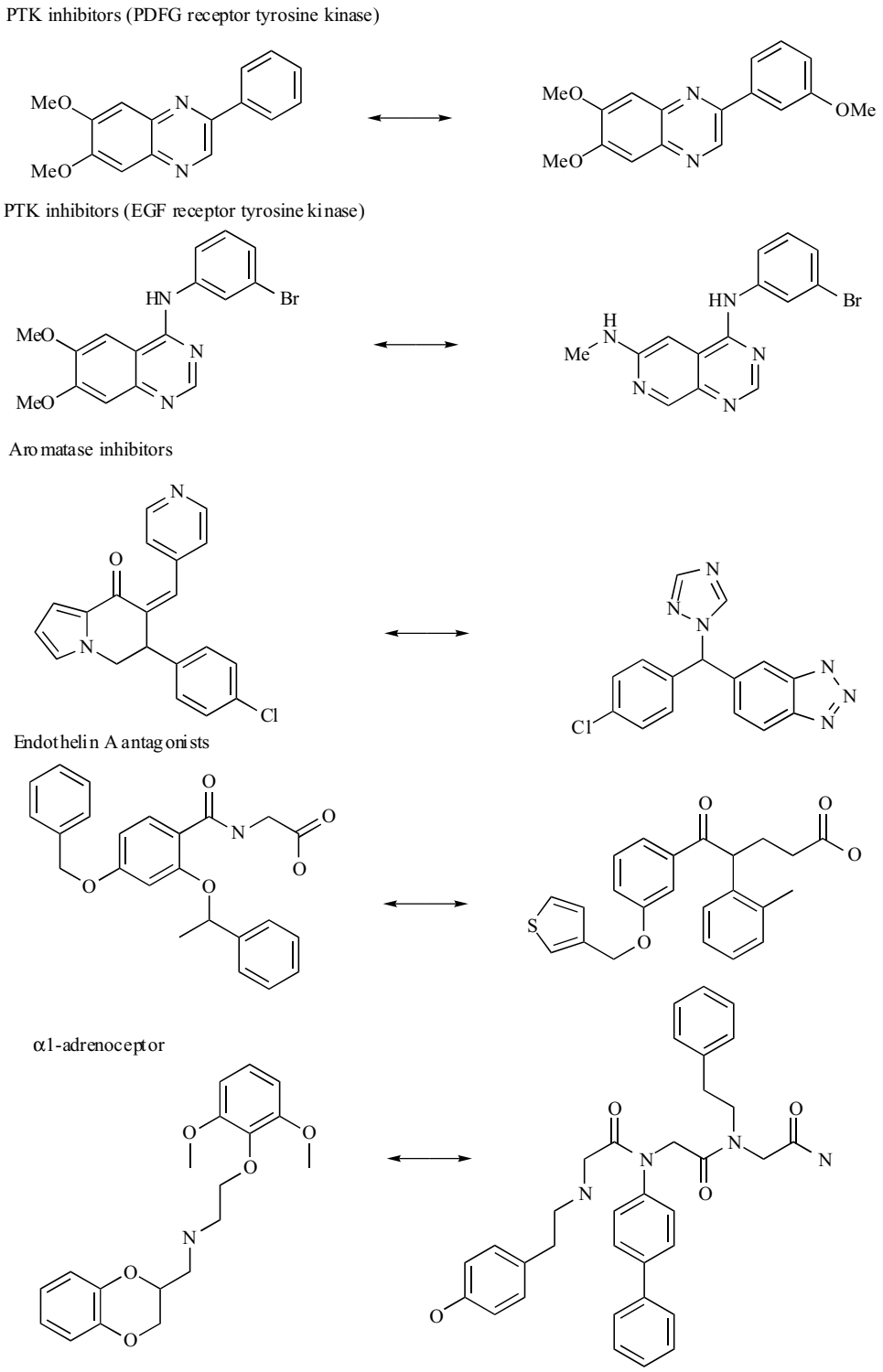

Fig. (1). Molecular similarity. Shown are examples of similarity relationships that have been correctly identified (in-house) using computational methods. From the top to the bottom, structures having similar activity become increasingly diverse. While molecular similarity can easily be "seen" in case of the first two examples, this is no longer possible for the other three remote relationships. Thus, computational similarity analysis must be capable of recognizing relatedness of molecules that sometimes goes far beyond the obvious and would not be predictable based on chemical intuition.

In computational chemistry and chemoinformatics, the need for extensive digital representations of chemical compounds and their physical properties has triggered the development of a wealth of different concepts over the years. Early approaches to defining descriptors focused on simple 2D drawings of molecules for applications in similarity searching [5]. These efforts were at least in part driven by the need to explore structure/property-activity relationships in compound databases, similar to the original application of substructure analysis [6]. Simple atom or bond counts and

substructures or structural keys [7] greatly accelerate substructure searching. Atom-pairs [8], topological torsions, and subsequent adaptations such as physiochemical descriptors [9] were developed. In addition, functional group descriptors, ring structures, or their hash codes were also found to be useful in describing features in molecules important for their biological activity [10-14]. Combinations of such descriptors can be encoded in keyed or hashed fingerprint representations for rapid molecular comparisons. These fingerprints are binary bit string representations of molecular structure and properties and often significantly vary in length and complexity [15]. These types of rather simple structural or property descriptors can be very helpful in understanding the results of similarity searching or compound classification as the interpretation directly relates to the understanding of molecular structure. In addition, various physical property descriptors, such as logP(o/w) can be conveniently calculated as a sum of fragment contributions. There are also a variety of topological indices [16], Zagreb [17], or Labute [18] descriptors that accumulate contributions from various molecular graphs, and the use of topological indices or related descriptors compresses information about molecules into a small number of quantities, thus permitting rapid comparison.

Table 1. Different Classes of Molecular Descriptors

1D
Molecular weight
Melting point
Number of halogen atoms
2D
Log of octanol/water part coefficient
Atomic connectivity index
Number of H-bond acceptors
Number of aromatic bonds
Number of rotatable bonds
3D
Solvent-accessible surface area
Van der Waals volume
3-point pharmacophore

Descriptor types are classified here according to the dimensionality of the molecular representation from which they are calculated.

Potential difficulties associated with the use of these representations result from their abundance. Chosen descriptor combinations might often describe compounds in too much detail and overemphasize molecular features that are not crucial for their biological activity. Simply put, using extensive digital representations, molecules with similar activity may often "look" rather different from each other. On the other hand, simplification of descriptors or reduction of their contributions might render molecules too similar and suggest false-positive structure-activity relationships. Thus, selecting the "right" set of descriptors for a search or classification problem, beyond educated guess or chemical intuition, continues to be a major obstacle in many cases. Only a rather limited number of investigation have thus far attempted to systematically rationalize descriptor selection through use of principal component analysis (PCA) [19,20], genetic algorithms [19,21], neural nets [20] or information theory [22].

In addition to 1D or 2D descriptors, many molecular descriptors are calculated from 3D structures. For example, three- or four-point pharmacophores [23] extend atom pairs or functional group representations, and descriptors based on solvent-exposed molecular surface area [24] or topomers [25] extend the concept of topological indices. Electronic properties calculated with semi-empirical methods and computationally simplified approaches such as transferable atom equivalents [26] allow the description of molecules at the level of electron distributions. A major problem involving the use of 3D descriptors is that they are typically calculated from hypothetical molecular conformations. There is of course no guarantee that such predictions indeed resemble bioactive conformations. In order to address these difficulties, methods of approximating 3D conformation-dependent properties such as the calculation of molecular surface area from 2D representation have recently been investigated [27,18]. In addition, pharmacophore fingerprints attempt to capture all possible spatial arrangements of pre-defined pharmacophore patterns in active molecules and compare them with others, thereby alleviating the need of predicting a specific conformation as the active one [28]. In the quantitative analysis of structure-activity relationships (QSAR), the conformational problem has been addressed by the introduction of 4D-QSAR [29], and the formalism has been further extended through inclusion of tautomers and charge states (i.e. 5D-QSAR) [30].

Although 3D descriptions of molecules should contain the maximum information achievable and thus be superior to 2D representations, the problems associated with predicting bioactive conformations and calculating conformation-dependent properties often eliminate principal advantages of 3D methods. In many cases, 2D and 3D descriptors and methods have yielded comparable performance in similarity analysis and compound classification and in others, either 2D or 3D approaches were found to be superior [31-34]. Clearly, performance depends to a large extent on the particular search or classification problem and, based on currently available data, it is hardly possible to assign general preference to either 2D or 3D approaches.

METRICS

The majority of diversity and classification techniques rely on pair-wise comparisons of molecules and a measure of distance between them. When discontinuous values for descriptors are used, as in binary fingerprints or pharmacophore counts, it is relatively simple to define a distance or similarity measure. In these cases, the Tanimoto, cosine, or dice distances (or coefficients) are among the preferred formulas [5]. Table **2** reports a number of these distance metrics. To quantitatively compare fingerprint overlap, the Tanimoto coefficient (Tc) has been, and continues to be, the most widely applied metric. It ranges from 0 to 1 for fingerprints having either no bits set on in common or identical bit settings, respectively. As has been

shown, possible Tc values are not evenly distributed and there are some statistical limitations to the use of the Tc metric [35, 36]. In addition, dependent on the specifics of the fingerprints, a Tc value of 1.0 does not necessarily mean that two molecules are identical.

Table 2. Conventional Similarity and Dissimilarity Functions

Similarity Metrics	
Tanimoto Coefficient:	$Tc = n_{ij} / (n_i + n_j - n_{ij})$
Dice Coefficient:	$Dc = 2n_{ij} / (n_i + n_j)$
Cosine Coefficient	$Cc = n_{ij} / (n_i\, n_j)^{1/2}$
Dissimilarity (Distance) Metrics	
Hamming Distance	$HDij = (n_i + n_j - 2n_{ij})$
Euclidean Distance	$EDij = (n_i + n_j - 2n_{ij})^{1/2}$
Average Distance	$D = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} D_{ij}}{n(n-1)}$

n_i and n_j are the number of bits set on (descriptor values) or descriptor values in fingerprints of molecules i and j, respectively, and n_{ij} is the number of bits or values in common to both molecules. D_{ij} is the distance between molecules i and j, D the average distance, and n the total number of molecules.

For continuous values of descriptors, other formulas can be applied. For example, the correlation coefficient between two sets of descriptors can be calculated and scaled between -1.0 and 1.0. Alternatively, scaling each descriptor to a unit range can be important for methods such as PCA. Comparisons of various distance metrics have recently been published [5]. In general, care must be taken not to generalize threshold values for the assessment of molecular similarity, since they typically depend on the classification method or similarity search tool that is applied. For example, a pioneering analysis of chemical neighborhood behavior implied that molecules reaching a Tanimoto similarity of 0.85 are often similar in activity [37]. However, as shown more recently, far too many potentially active compounds were eliminated from a large compound collection when this similarity threshold value was applied in search calculations [38].

For diversity calculations and compound classification, it is important to consider that the choice of descriptors, their relative weighting, the similarity or distance metrics, and the selected threshold values may have a substantial effects on the chemical reasonableness of the results. In many instances, the application of distance measures produces a high-dimensional problem (N descriptors by the square of the number of compounds), which ultimately limits the number of molecules that can be effectively processed.

METHODS

At the heart of compound classification or diversity analysis are approaches to organize compounds into groups of similar ones or spread them out into distinct areas of chemical space, respectively. Compounds can be classified according to any chosen compound characteristic or property. For example, compound classification techniques can group compounds according to biological activity [15], differentiate drugs from non-drugs [39,40], or systematically distinguish between molecules from synthetic or natural sources [41].

How is the predictive value of these methods evaluated? If we take activity as an example, the measurement of the quality of a classification system is rather clear. One would like to correctly identify all actives in a collection of molecules while maintaining a low false-positive rate. Typically, a selection of active and inactive compounds would be combined as a training set for a particular method and a collection of active compounds reserved as a prediction set. The rate at which active molecules are identified is compared to random selection and expressed as an enrichment factor. Other measures of predictive performance or enrichment have also been proposed such as the number of compounds that need to be tested until half the number of actives are found (A50) [9]. For the selection of representative subsets from existing compound collections, measures of success are usually more subjective because, as discussed above, any solution to the problem is much influenced by the design of chemical reference space and the descriptors applied.

Dissimilarity and Diversity Analysis

In the context of molecular diversity, methods for dissimilarity analysis and diversity design should be distinguished, although their boundaries are rather fluid in some cases. Dissimilarity-based methods aim at selecting subsets of compounds from larger collections that are most dissimilar from each other [2,42]. Different algorithms have been developed for dissimilarity selection [42], all of which have in common that they rely on pair-wise comparison of compound distances in descriptor spaces. The computational complexity of maximum dissimilarity methods is on the order of $O(kn)$ to $O(k^2n)$, with k being the size of the subset and n the size of the original collection. Popular methods include the OptiSim [43] and DivPick [44] algorithms. In DivPick, which is among the simplest approaches, compounds are randomly selected and only those are obtained that are sufficiently diverse from previous selections based on a user-defined cutoff value of descriptor distances. This technique was recently used in a combination with hierarchical clustering steps to classify compound sets based on dissimilarity [36,45]. Computationally more efficient techniques have been described including the centroid-based diversity sorting algorithm [46] having a complexity on the order of $O(n)$.

Recently designed algorithms for diversity estimation have attempted to circumvent the problem of pair-wise distance comparisons altogether by probability sampling [47,48]. These types of methods essentially establish a link between dissimilarity and diversity approaches. In diversity design, the major task is to place compounds in chemical space far enough apart so that some degree of uniform coverage is achieved without overpopulating subspaces and creating redundancy [3,49]. It is intuitively clear that such strategies have played a crucial role in the design of diverse combinatorial libraries beginning in the mid 1990s. These

"inductive" diversity estimations can be carried out either at the level of reagents or reaction products [50] and taking into account 2D or 3D properties including pharmacophore models [51]. Reagent-based diversity selection is extremely fast but not as accurate as product-based evaluation for some descriptor combinations. However, diversity analysis at the level of products becomes quickly prohibitive if full enumeration of compound libraries is required. However, if probability sampling of reagents is applied, the diversity distribution of the resulting library can be estimated without the need to fully enumerate the products [47], presenting a significant advancement for efficient library design.

QSAR

Among classification techniques, QSAR [52] and cluster analysis [3] have perhaps the longest history in computational medicinal chemistry. The underlying assumption of classical QSAR analysis is that sets of compounds with differential activity follow linear structure/property-activity relationships, which is an approximation in many cases. Original applications of QSAR based on linear free energy relationships have been extended to all types of 2D descriptors, 3D fields [53], and multi-dimensional formalisms, as mentioned above [29,30]. Statistical measures of success are relatively easy to determine but, at least for classical QSAR, the prediction range is often quite narrow. For analyzing large data sets such as high-throughput screening (HTS) data where biological response results from a single compound dose, a binary QSAR formalism has been introduced [54]. Based on a chosen activity threshold criterion, test compounds are simply classified as, for example, active or inactive, and the method uses Bayesian probabilities and probability functions to estimate the likelihood that other compounds are also active.

Clustering and Decision Tree Techniques

Conventional clustering methods are divided into hierarchical and non-hierarchical techniques. Figure **2** illustrates different clustering techniques. Hierarchical methods build relationships between clusters and are further distinguished as divisive or agglomerative clustering, dependent on whether the process starts from a large clusters or single compounds, respectively. In agglomerative clustering, pair-wise similarity relationships are established between molecules and clusters that are overall most similar to each other are combined. By contrast, divisive clustering starts from a large cluster and further divides it (and the resulting clusters) based on lowest pair-wise compound similarities. Non-hierarchical clustering algorithms group molecules together into a pre-defined number of clusters without analyzing inter-cluster relationships. For chemical classification with non-hierarchical methods, Jarvis-Patrick (JP) clustering [55] was shown to be the preferred approach [56]. JP clustering is based on nearest neighbor analysis in descriptor space. Two molecules are included in the same

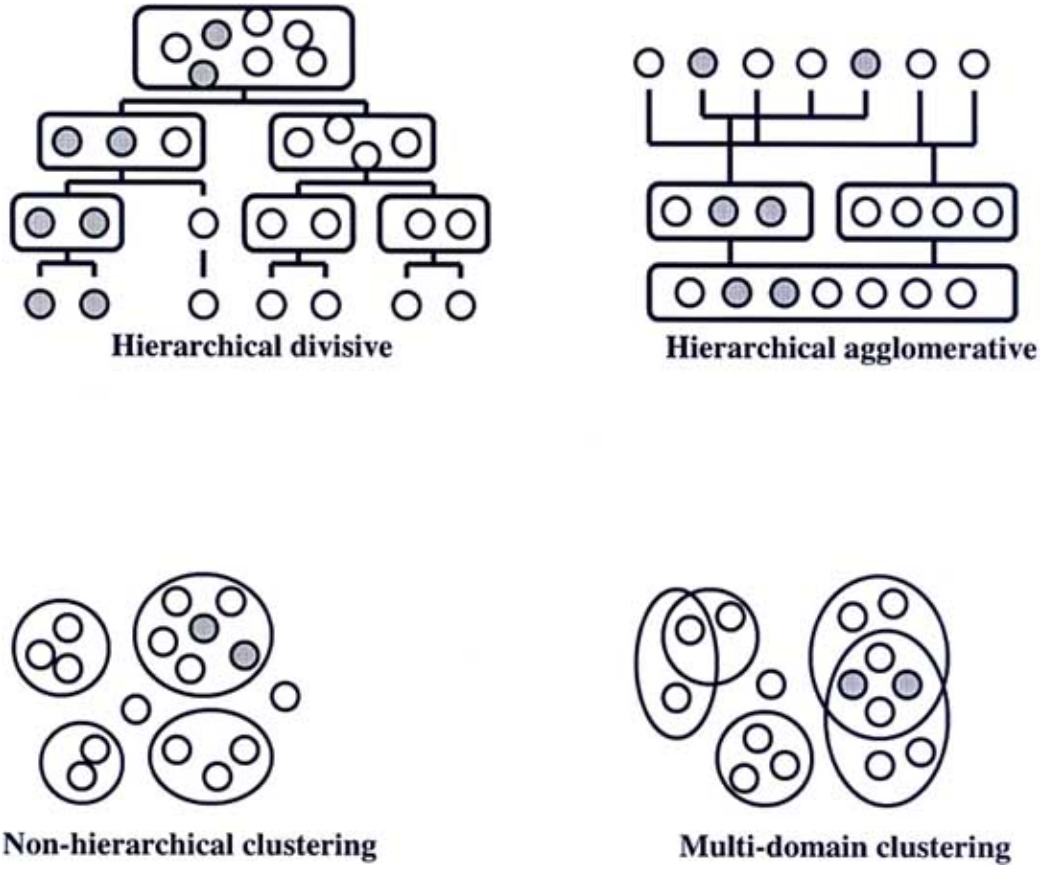

Fig (2). Clustering techniques. Different clustering methods (as discussed in the text) are schematically illustrated. Gray dots represent compounds having similar activity.

cluster if they have a minimum number of nearest neighbors. Among hierarchical methods, (agglomerative) Ward clustering [57] was found to be most effective (and better than JP) [58].

Related to hierarchical clustering are methods that classify compounds along decision trees such as recursive partitioning (RP) [59,60]. At each node of the tree, RP utilizes two-state descriptors (such as structural fragments) to divide data sets into subsets of molecules that share or do not share these descriptors, and the average biological activity of these subsets is calculated. Ultimately, this procedure leads to an enrichment of active molecules at terminal nodes where active structures are associated with final descriptor settings. The obtained rules can then be applied to search databases for active compounds. Different from clustering methods, RP has low complexity and can be applied to very large descriptor pools and compound data sets. However, the availability of many different descriptors increases the risk that classification results might be easily trapped in narrow regions of descriptor space. Therefore, RP has recently been combined with simulated annealing techniques [61]. Such stochastic solutions to descriptor selection make use of an analogy to physical systems by applying the Metropolis criterion to the fitness function. Here less than optimal choices of descriptors may initially be selected in order to overcome barriers between local minima and the global optimum of the classification or diversity scheme.

Conventional clustering methods and RP have in common that active compounds can only belong to one final subset. However, this can be problematic if compounds having similar activity are structurally diverse and separated during the clustering or partitioning process. In order to address these shortcomings, techniques such as fuzzy [44,62] and multi-domain [63] clustering have recently been introduced that permit active molecules to occur in more than one final group and make it possible to establish overlapping and more detailed structure-activity relationships. Multi-domain clustering also utilizes a decision tree structure, similar to RP, but incorporates other components such as simulated annealing and genetic algorithms in the clustering process.

Partitioning and Cell-Based Methods

In contrast to cluster methodologies, compound partitioning does not depend on pair-wise compound and distance comparisons and is therefore in principle applicable to much larger data sets. In partitioning, a coordinate system must be generated in chemical space that defines where each compound maps based on its calculated descriptor values. The computational efficiency of partitioning methods is much influenced by the dimensionality of the chemical reference space and the complexity of operations required for its manipulation and segmentation [64]. Partitions in multi-dimensional space are generated by dividing descriptor axes into regularly spaced intervals (a process called binning). The generation of evenly spaced and populated partitions or cells for diversity selection or compound classification becomes more difficult with increasing dimensionality of chemical space and, consequently, low-dimensional space representations are often preferred.

Cell-based partitioning in low-dimensional chemistry spaces became a popular approach following the introduction of the BCUT metric [65]. BCUTs are uncorrelated composite molecular descriptors that combine information about molecular connectivity, inter-atomic distances, and diverse molecular properties. Their application makes it possible to construct reference spaces that typically consist of six orthogonal axes [65]. For visualization purposes, the dimensionality can often be further reduced to obtain 3D representations. This is achieved by identifying the BCUT axes around which the majority of active compounds are located and eliminating the ones that do not contribute to these distributions [66].

Dimension reduction of high-dimensional descriptor spaces for partitioning has also been accomplished by several methods including PCA [67], single value decomposition [68], including multi-dimensional scaling [69], or non-linear mapping [70]. Techniques like single value decomposition and PCA reduce the dimensionality of chemical spaces by eliminating descriptor correlation effects and generating new orthogonal descriptor vectors. For example, principal components consist of linear combinations of the original descriptors assigning high weight to those that are most important for capturing the variability within the compound data set under investigation. Multi-dimensional scaling and non-linear mapping are closely related techniques that attempt to find those variables that preserve original data distributions in lower dimensions. However, these methods require distance comparisons, which limit the size of the compound data sets, similar to conventional clustering. As has been shown, such limitations might be circumvented when these approaches are combined with probability sampling and machine learning techniques in order to estimate global data distributions from smaller reference sets [70].

Other recent studies have demonstrated that effective compound partitioning does not necessarily depend on the generation of low-dimensional descriptor spaces. The median partitioning (MP) algorithm makes use of statistical medians of descriptor value distributions, transforms numerical descriptors into a binary classification scheme, and divides molecular data sets in subsequent steps into an increasing number of unique partitions (n descriptors produce a total of 2^n partitions) [71]. As illustrated in (Fig. **3**), MP operates in high-dimensional space, as opposed to cell-based methods. MP has been shown to be an effective approach for diversity selection [71] and also classification of bioactive compounds [72], thereby achieving accurate results in up to 20-dimensional reference spaces.

Neural Networks

Neural networks (NN) and self-organizing maps are also used to develop complex models for compound classification [73]. NNs are arrays of connected mathematical models that are organized in different layers and simulate adaptive learning processes. They are trained to distinguish between different objects and their properties in learning sets and used to make predictions. In self-organizing maps, multi-dimensional descriptor space is reduced to a 2D representation with toroidal shape, and these maps can be used as a clustering tool for molecular data sets [74]. NN

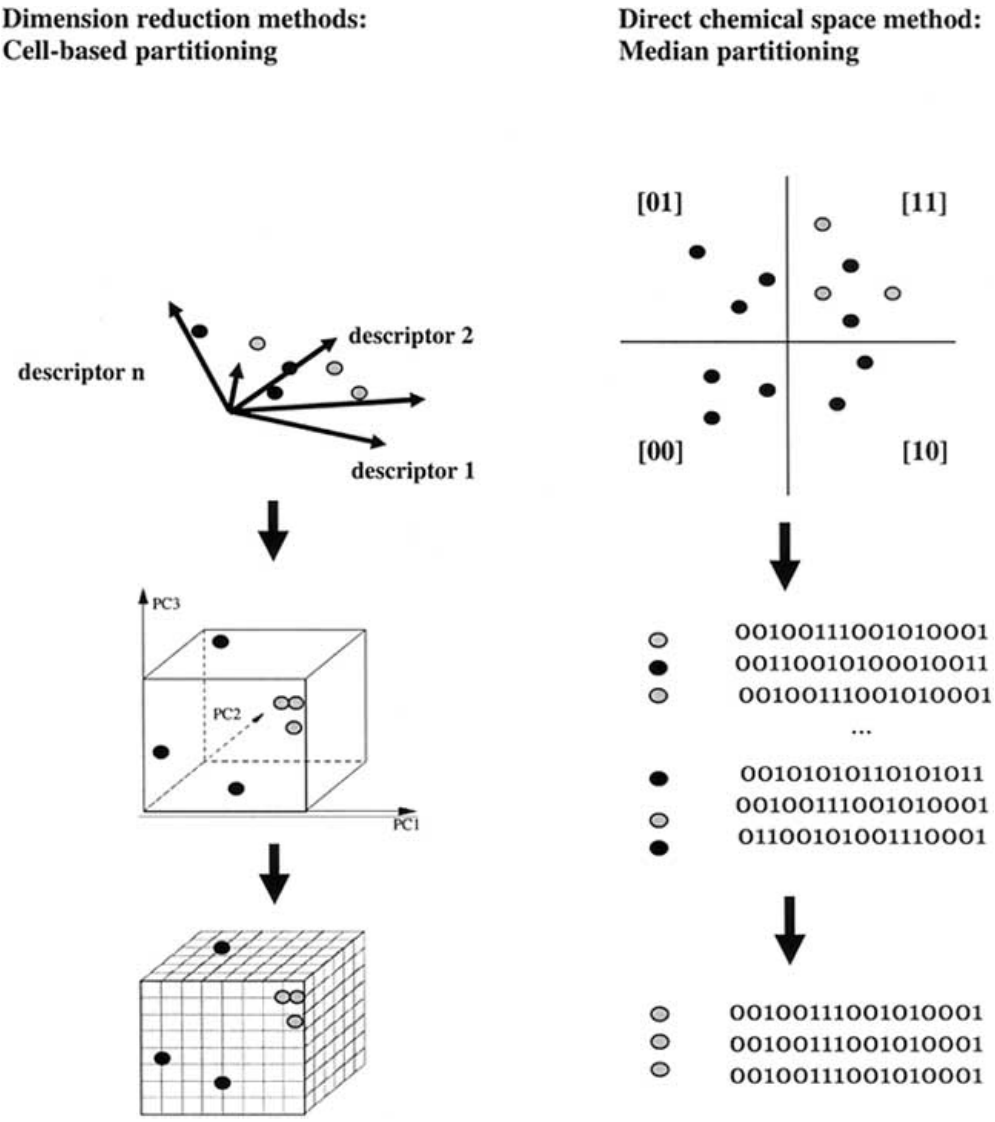

Fig (3). Partitioning algorithms. The figure highlights differences between cell-based partitioning, which involves dimension reduction and orthogonalization of descriptor space, and direct space partitioning. Dimension reduction is achieved in this case by deriving uncorrelated composite descriptors from original descriptor settings, and binning produces cells for partitioning in low-dimensional space. By contrast, in the type of direct space partitioning shown here, n numerical descriptors are transformed into a binary classification scheme in subsequent steps, thereby producing 2^n partitions, each of which is characterized by a unique partition code. Gray dots represent a class of active compounds and black dots other database compounds.

models are usually extremely complex and difficult to interpret but are able to derive relationships between chemical and biological data that would not easily be achieved by simpler functions. Importantly, different from QSAR-analysis, NN models can capture non-linear structure/property-activity relationships. Self organizing maps are among preferred techniques for unsupervised machine learning, whereas back propagating NNs are often also used for supervised learning [73]. Some systematic differentiation between drugs and non-drugs was first accomplished by NN simulations [39,40].

Other Techniques

Genetic algorithms have been successfully applied to aid in diversity analysis and compound classification, especially for systematic descriptor selection [19,21]. Following this approach, descriptors, weighting schemes, and/or calculation parameters are encoded as a binary "chromosome" and each bit setting provides a combination of these variables for classification calculations. Thus, the genetic algorithm is typically coupled to a classification method and results are evaluated using a fitness function. Chromosomes producing best intermediate solutions are subjected to mutation and cross-over operations, and the process is continued until a convergence criterion of the fitness function is reached.

Trend vectors are calculated as the first moment of activity in descriptor space and are a very fast technique to evaluate descriptor importance relative to biological activity [8]. Descriptors for each compound are weighted by the distance of each molecule from the average activity in a

training set. Trend vector calculations are equivalent to partial least squares analysis and the technique can be applied to binary or continuous activity data [75].

Support vector machine represents a classification technique reminiscent of linear discriminant analysis [76]. It divides compounds along a complex hyperplane in descriptor space so that the distance from active and inactive compounds to the plane is minimized and, ultimately, arrives at a binary formulation of activity (i.e. active versus inactive). In an iterative fashion, compounds can be tested and added to the model in order to increase its information content. Test compounds are scored and considered active if the score exceeds a pre-defined threshold value, similar to the binary QSAR approach discussed above.

Binary kernel discrimination has recently been applied to classification of data sets consisting of active and inactive compounds [77]. The method is non-parametric and uses binary descriptors or fingerprints to estimate the probability of a molecule to be active. For two molecules under comparison fingerprints are calculated and the number of bit positions that differ is determined and used as input for a kernel function. The function is summed for all active and inactive compounds in a training set and a single parameter of the kernel function is varied to obtain the best relative ranking of actives. This parameter is then applied to calculate kernel functions for test compounds and predict active compounds based on their ranking. The method has drawbacks in that its complexity scales quadratically with the number of molecules in the training set and that training set calculations must be repeated several times for parameter optimization.

APPLICATIONS

Molecular diversity and compound classification techniques are widely used in drug discovery settings. It is fair to say that none of the methods discussed here is currently accepted as a gold standard in the field (and considered generally superior to others). In fact, many institutions or companies develop their own methodologies, and the way different techniques are applied varies in part significantly. Compound classification techniques have also been adapted for virtual screening of large compound databases. An increasing number of interesting and successful applications are reported in the literature and, in order to complement the description of methods, some representative examples will be discussed in the following.

Diversity Analysis

In the pharmaceutical industry, dissimilarity-based methods and selection schemes have been intensely applied in recent years to design and implement compound acquisition strategies. A number of algorithms have been optimized to add diverse compounds from external libraries to in-house collections [78,79]. For example, LASSOO [79] uses structural key-type fingerprints to characterize members of an existing library and exploit differences in the distribution of fingerprints settings in order to identify novel dissimilar compounds. Many dissimilarity or clustering methods applied to compound acquisition can only compare data sets of limited size to existing libraries, due to pair-wise molecular comparisons involved. Only recently, the MP method was introduced that makes it possible to efficiently select diverse subsets from compound sources containing several million molecules [71]. The algorithm is capable of doing so because it does not rely on pair-wise compound comparisons and because it directly operates in multi-dimensional chemical space (and does not require complex mathematical space transformations).

While achieving chemical diversity of compound collections is still an important goal, it is no longer the principal objective, as it has been recognized that diversity must be complemented, and sometimes constrained, by desired drug-like properties (although it remains difficult to understand, and predict, what exactly renders a molecule drug-like) [80,81]. These insights have influenced both compound acquisition and library design strategies. Efforts to optimize chemical diversity of combinatorial libraries by reagent- or product-based design are well documented in the literature [82-84]. Furthermore, scalable techniques that circumvent explicit enumeration and compound storage have been introduced for the generation of diverse *in silico* virtual libraries containing literally hundreds of millions of molecules [85]. However, a recent trend in this field is to build libraries that are not maximally diverse but constrained to follow desired property profiles. This task can be quite challenging because, in these cases, library design requires the parallel optimization of multiple and sometimes conflicting parameters. Therefore, multi-objective optimization and design techniques have recently been introduced [86-88]. For example, Gillet *et al*. have applied a genetic algorithm to generate diverse libraries with drug-like properties [88]. Different combinatorial libraries were designed and described using MoSELECT. This program suggests equally plausible solutions that compromise between diversity and property objectives from which the library designer can choose. Instead of applying a weighted-sum fitness function for every property to be optimized, MoSELECT is based on a multi-objective evolutionary framework, called MOGA, where the ranking of parameter populations is based on dominance instead of standard fitness functions [89].

Compound Classification

Among the classification methods that have recently experienced significant interest in drug discovery, are cell-based partitioning and decision tree methods (first and foremost, RP). The popularity of cell-based partitioning is at least in part due to its conceptual elegance and the attractiveness of decision tree techniques is to a large extent due to its extreme computational efficiency. The predictive value of these approaches has been demonstrated in a number of case studies. For example, the BCUT metric has been successfully applied in library comparison [90], classification of enzyme inhibitors [91], and QSAR-like analysis [92]. Similarly, PCA-based partitioning has yielded high accuracy in classifying compound belonging to diverse biological activity classes [67]. Recursive partitioning is being widely applied to distinguish between active and inactive compounds and build predictive models from HTS data sets, as further discussed below. In these investigations, enrichment factors of greater than 10-fold in classifying

active compounds are frequently achieved as well as 5-fold or greater improvements of hit rates in screens [59,60,93].

In addition to these popular approaches, recently developed methods are beginning to be successfully applied. For example, multi-domain clustering could distinguish between many anti-HIV compounds and inactive molecules and associate different chemotypes with anti-HIV activity [63,94]. Using binary kernel discrimination, several fold improvements in hit rates were obtained [77], similar to RP. In a recent study, binary kernel discrimination outperformed other *in silico* screening techniques, including substructure and fingerprint searching and trend vector analysis, whereas support vector machines performed worse than any of these methods [95]. Thus, it can be expected that methods such as multi-domain clustering and binary kernel discrimination will be more extensively applied in the near future.

Another noteworthy trend is the increasing application of intuitive and sometimes rather simple but useful filter functions to process very large number of compounds and enrich libraries with compound having desired properties [96]. For example, a filter monitoring preferred functional groups in drugs recognized drug-like compounds with 30% higher frequency than other database molecules [97]. In addition, property profiles of drug-like molecules have been generated using a number of simple molecular descriptors that display some degree of drug-selectivity in database analysis [98]. These similar efforts demonstrate that focusing or enrichment of compound libraries does not necessarily require the application of complex computational models such as neural nets, provided statistical criteria or knowledge-based approaches can be applied.

Virtual Screening

More or less all compound classification methods discussed above have been adapted for virtual screening (and complement similarity search and docking techniques) [99]. Some reports are rather encouraging. For example, in benchmark calculations, a NN-based virtual screen identified almost 90% of all cytochrome P450 3A4 inhibitors in a compound database [100], and an iterative virtual screen based on hierarchical clustering produced hit rates between 25% and 39% for an anti-HIV data [101]. Furthermore, applying recursive median partitioning, an adaptation of the MP algorithm for virtual screening of very large databases, more than 1000-fold hit enrichment over random compound selection was achieved for several classes of inhibitors and antagonists [102].

An interesting development in virtual screening is the formation of interfaces to HTS [99]. Although the costs per well or compound for large magnitude screening campaigns have become lower over the years, the number of compounds in screening data sets has increased substantially. Consequently, there are often budgetary reasons for limiting the size of screening sets, and this is increasingly being attempted by chemoinformatics analysis. Furthermore, in the virtual screening community, evidence is accumulating that the efficiency of screening programs can be increased through computational support. The interplay between virtual screening and HTS involves different components. Screening sets can be focused by compound classification or filtering. In addition, methods such as binary QSAR, RP, or binary kernel discrimination, as discussed above, are used to analyze HTS data sets and build predictive models of biological activity, which can in turn be applied for virtual screening or focusing of source libraries [99]. Perhaps the most promising development in this area aims at combining virtual and wet screening in an iterative manner, a process termed sequential screening [99,103]. For example, in a benchmark study on various anti-cancer compounds, two-stage sequential screening produced hit rates of up to 40%, and an analysis of several such experiments revealed that two iterations combining virtual and wet screening were typically sufficient to identify more than 50% of all hits by testing of only about 10% of compounds in screening libraries [103]. Such findings illustrate the potential of sequential screening or related concepts.

CONCLUSIONS

Methods to assess molecular diversity and to classify compounds have been discussed. In drug discovery research, these approaches have become popular over the past 15 years or so. Although the concept of molecular diversity is beginning to be re-evaluated in the context of molecular property analysis and the study of drug-likeness, there is little doubt that similarity, dissimilarity, and diversity analysis will continue to play an important role at the interface between chemoinformatics and synthetic and medicinal chemistry. In addition, many compound classification methods are under continuous development. Clustering and partitioning algorithms have been adapted and refined for increasingly demanding applications such as effective screening of very large databases containing millions of molecules. Moreover, new concepts for the identification of active compounds have recently been introduced. We anticipate that virtual and high-throughput screening programs will be more integrated in the future than they are today. In addition, the development of advanced compound filter functions, predictive property profiles, and novel classification methods continues to experience an increase in interest in the chemoinformatics community.

REFERENCES

[1] Johnson, M. A.; Maggiora, G. M. *Concept and Applications of Molecular Similarity;* Wiley: New York: **1990**.

[2] Lajiness, M. Molecular Similarity-Based Methods for Selecting Compounds for Screening., in *Computational Chemical Graph Theory*, Rouvray, D. H., editor; Nova Science Publishers: New York, **1990**; pp. 299-316.

[3] Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. *J. Med. Chem.*, **1995**, *38*, 1431.

[4] Willett, P. *Similarity and Clustering in Chemical Information Systems;* Research Studies Press: Letchworth: **1987**.

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

[6] Cramer, R. D.; Redl, G.; Berkoff, C. E. *J. Med. Chem.*, **1974**, *17*, 533.

[7] Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. *J. Chem. Inf. Comput. Sci.*, **2002**, *42*, 1273.

[8] Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. *J. Chem. Inf. Comput. Sci.*, **1985**, *25*, 64.

[9] Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose J.D.; Mosley, R. T.; Sheridan, R. P. *J. Chem. Inf. Comput. Sci.*, **1996**, *36*, 118.

[10] Ertl, P. *J. Chem. Inf. Comput. Sci.*, **2003**, *43*, 374.

[11] Bemis, G. W.; Murcko, M. A. *J. Med. Chem.*, **1996**, *39*, 2887.
[12] Chen, X.; Reynolds, C. H. *J. Chem. Inf. Comput. Sci.*, **2002**, *42*, 1407.
[13] Nilakantan, R.; Bauman, N.; Haraki, K. S.; Venkataraghavan, R. *J. Chem. Inf. Comput. Sci.*, **1990**, *30*, 65.
[14] Xu, J. *J. Med. Chem.*, **2002**, *45*, 5311.
[15] Bajorath, J. *J. Chem. Inf. Comput. Sci.*, **2001**, *41*, 233.
[16] Hall, L. H.; Kier, L. B.; Brown, B. B. *J. Chem. Inf. Comput. Sci.*, **1995**, *35*, 1074.
[17] Basak, S. C.; Balaban, A. T.; Grunwald, G. D.; Gute, B. D. *J. Chem. Inf. Comput. Sci.*, **2000**, *40*, 891.
[18] Labute, P. *J. Mol. Graph. Model.*, **2000**, *18*, 464.
[19] Xue, L.; Bajorath, J. *J. Chem. Inf. Comput. Sci.*, **2000**, *40*, 801.
[20] Brustle, M.; Beck, B.; Schindler, T.; King, W.; Mitchell, T.; Clark, T. *J. Med. Chem.*, **2002**, *45*, 3345.
[21] Gillet, V. J.; Willett, P.; Bradshaw, J. *J. Chem. Inf. Comput. Sci.*, **1998**, *38*, 165.
[22] Godden, J. W.; Bajorath, J. *J. Chem. Inf. Comput. Sci.*, **2002**, *42*, 87.
[23] Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. *J. Med. Chem.*, **1999**, *42*, 3251.
[24] Clark, D. E. *J. Pharmaceutical Sci.*, **1999**, *88*, 807.
[25] Cramer, R. D.; Clark, R. D.; Patterson, D. E.; Ferguson, A. M. *J. Med. Chem.*, **1996**, *39*, 3060.
[26] Song, M.; Breneman, C. M.; Bi, J.; Sukumar, N.; Bennett, K. P.; Cramer, S.; Tugcu, N. *J. Chem. Inf. Comput. Sci.*, **2002**, *42*, 1347.
[27] Ertl, P.; Rohde, B.; Selzer, P. *J. Med. Chem.*, **2000**, *43*, 3714.
[28] Mason, J. S.; Cheney, D. L. *Pac. Symp. Biocomput.*, **2000**, *5*, 576.
[29] Duca, J. S.; Hopfinger, A. J. *J. Chem. Inf. Comput. Sci.*, **2001**, *41*, 1367.
[30] Vedani, A.; Dobler, M. *J. Med. Chem.*, **2002**, *45*, 2139.
[31] Sheridan, R. P.; Miller, M. D.; Underwood, D. J.; Kearsley, S. K. *J. Chem. Inf. Comput. Sci.*, **1996**, *36*, 128.
[32] Brown, R. D.; Martin, Y. C. *J. Chem. Inf. Comput. Sci.*, **1997**, *37*, 1.
[33] Matter, H. *J. Med. Chem.*, **1997**, *40*, 1219.
[34] Xue, L.; Stahura, F. L.; Godden, J. W.; Bajorath, J. *J. Chem. Inf. Comput. Sci.*, **2001**, *41*, 394.
[35] Flower, D. R. *J. Chem. Inf. Comput. Sci.*, **1998**, *38*, 379.
[36] Reynolds, C. H.; Druker, R. *J. Chem. Inf. Comput. Sci.*, **1998**, *38*, 305.
[37] Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. *J. Med. Chem.*, **1996**, *39*, 3049.
[38] Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. *J. Med. Chem.*, **2002**, *45*, 4350.
[39] Ajay, A.; Walters, W. P.; Murcko, M. A. *J. Med. Chem.*, **1998**, *41*, 3314.
[40] Sadowski, J.; Kubinyi, H. *J. Med. Chem.*, **1998**, *41*, 3325.
[41] Stahura, F. L.; Godden, J. W.; Xue, L.; Bajorath, J. *J. Chem. Inf. Comput. Sci.*, **2000**, *40*, 1245.
[42] Willett, P. *J. Comput. Biol.*, **1999**, *6*, 447.
[43] Clark, R. D. *J. Chem. Inf. Comput. Sci.*, **1997**, *37*, 1181.
[44] Nilakantan, R.; Bauman, N.; Haraki, K. S. *J. Comput. Aided Mol. Des.*, **1997**, *11*, 447.
[45] Reynolds, C. H.; Tropsha, A.; Pfahler, L. B.; Druker, R.; Chakravorty, S.; Ethiraj, G.; Zheng, W. *J. Chem. Inf. Comput. Sci.*, **2001**, *41*, 1470.
[46] Holliday, J. D.; Ranade, S. S.; Willett, P. A. *Quant. Struct. -Act. Relat.*, **1995**, *14*, 501.
[47] Beroza, P.; Bradley, E. K.; Eksterowicz, J. E.; Feinstein, R.; Greene, J.; Grootenhuis, P. D.; Henne, R. M.; Mount, J.; Shirley, W. A.; Smellie, A.; Stanton, R. V.; Spellmeyer, D. C. *J. Mol. Graph. Model.*, **2000**, *18*, 335.
[48] Agrafiotis, D. K. *J. Chem. Inf. Comput. Sci.*, **2001**, *41*, 159.
[49] Martin, Y. C. *J. Comb. Chem.*, **2001**, *3*, 231.
[50] Jamois, E. A.; Hassan, M.; Waldman, M. *J. Chem. Inf. Comput. Sci.*, **2000**, *40*, 63.
[51] Martin, E. J.; Hoeffel, T. J. *J. Mol. Graph. Model.*, **2000**, *18*, 383.
[52] Hansch, C.; Hoekman, D.; Leo, A.; Weininger, D.; Selassie, C. D. *Chem. Rev.*, **2002**, *102*, 783.
[53] Cramer, R. D.; Patterson, D. E.; Bunce, J. D. *J. Am. Chem. Soc.*, **1988**, *110*, 5959.
[54] Labute, P. *Pac. Symp. Biocomput.*, **1999**, *4*, 444.
[55] Jarvis, R. A.; Patrick, E. A. Clustering using a similarity measure based on shared near neighbors, in *IEEE Trans. Comput.,* **1973**; Chapter C-22, pp. 1025-1034.
[56] Willett, P.; Wintermann, V.; Bawden, D. J. *J. Chem. Inf. Comput. Sci.*, **1986**, *26*, 109.
[57] Ward, J. H. *J. Am. Stat. Assoc.*, **1963**, *58*, 236.
[58] Brown, R. D.; Martin, Y. C. *J. Chem. Inf. Comput. Sci.*, **1996**, *36*, 572.
[59] Chen, X.; Rusinko, A. I.; Young, S. S. *J. Chem. Inf. Comput. Sci.*, **1998**, *38*, 1054.
[60] Rusinko, A., III; Farmen, M. W.; Lambert, C. G.; Brown, P. L.; Young, S. S. *J. Chem. Inf. Comput. Sci.*, **1999**, *39*, 1017.
[61] Blower, P.; Fligner, M.; Verducci, J.; Bjoraker, J. *J. Chem. Inf. Comput. Sci.*, **2002**, *42*, 393.
[62] Feher, M.; Schmidt, J. M. *J. Chem. Inf. Comput. Sci.*, **2003**, *43*, 810.
[63] Tamura, S. Y.; Bacha, P. A.; Gruver, H. S.; Nutt, R. F. *J. Med. Chem.*, **2002**, *45*, 3082.
[64] Stahura, F. L.; Bajorath, J. *Curr. Med. Chem.*, **2003**, *10*, 707.
[65] Pearlman, R. S.; Smith, K. M. *Perspect. Drug Discov. Design*, **1998**, *9*, 339.
[66] Pearlman, R. S.; Smith, K. M. *J. Chem. Inf. Comput. Sci.*, **1999**, *39*, 28.
[67] Xue, L.; Bajorath, J. *J. Chem. Inf. Comput. Sci.*, **2002**, *42*, 757.
[68] Xie, D.; Tropsha, A.; Schlick, T. *J. Chem. Inf. Comput. Sci.*, **2000**, *40*, 167.
[69] Agrafiotis, D. K.; Lobanov, V. S. *J. Comput. Chem.*, **2001**, *22*, 1712.
[70] Agrafiotis, D. K.; Lobanov, V. S. *J. Chem. Inf. Comput. Sci.*, **2000**, *40*, 1356.
[71] Godden, J. W.; Xue, L.; Kitchen, D. B.; Stahura, F. L.; Schermerhorn, E. J.; Bajorath, J. *J. Chem. Inf. Comput. Sci.*, **2002**, *42*, 885.
[72] Godden, J. W.; Xue, L.; Bajorath, J. *J. Chem. Inf. Comput. Sci.*, **2002**, *42*, 1263.
[73] Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design;* 2nd. ed. Wiley-VCH ed.; Weinheim, **1999**.
[74] Rabow, A. A.; Shoemaker, R. H.; Sausville, E. A.; Covell, D. G. *J. Med. Chem.*, **2002**, *45*, 818.
[75] Sheridan, R. P.; Kearsley, S. K. *Drug Discov. Today*, **2002**, *7*, 903.
[76] Burbridge, R.; Trotter, M.; Buxton, B.; Holden, S. *Comput. Chem.*, **2001**, *26*, 5.
[77] Harper, G.; Bradshaw, J.; Gittins, J. C.; Green, D. V.; Leach, A. R. *J. Chem. Inf. Comput. Sci.*, **2001**, *41*, 1295.
[78] Turner D.B.; Tyrrell S.M.; Willett, P. *J. Chem. Inf. Comput. Sci.*, **1996**.
[79] Koehler, R. T.; Dixon, S. L.; Villar, H. O. *J. Med. Chem.*, **1999**, *42*, 4695.
[80] Manly, C. J.; Louise-May, S.; Hammer, J. D. *Drug Discov. Today*, **2001**, *6*, 1101.
[81] Clark, D. E.; Pickett, S. D. *Drug Discov. Today*, **2000**, *5*, 49.
[82] Gorse, D.; Lahana, R. *Curr. Opin. Chem. Biol.*, **2000**, *4*, 287.
[83] Agrafiotis, D. K.; Lobanov, V. S.; Salemme, F. R. *Nat. Rev. Drug Discov.*, **2002**, *1*, 337.
[84] Gillet, V. J. *J. Comput. Aided Mol. Des*., **2002**, *16*, 371.
[85] Lobanov, V. S.; Agrafiotis, D. K. *Comb. Chem. High Throughput Screen.*, **2002**, *5*, 167.
[86] Agrafiotis, D. K. *J. Comput. Aided Mol. Des*., **2002**, *16*, 335.
[87] Brown, R. D.; Hassan, M.; Waldman, M. *J. Mol. Graph. Model.*, **2000**, *18*, 427, 537.
[88] Gillet, V. J.; Khatib, W.; Willett, P.; Fleming, P. J.; Green, D. V. *J. Chem. Inf. Comput. Sci.*, **2002**, *42*, 375.
[89] Fonseca C.M.; Fleming, P. J. Genetic algorithms for multiobjective optimization: formulation, discussion and generalisation, in *Genetic Algorithms: Proceedings of the fifth International Conference*, Forrest, S. M. K., editor; San Mateo,CA, **1993**; pp. 416-423.
[90] Schnur, D. *J. Chem. Inf. Comput. Sci.*, **1999**, *39*, 36.
[91] Pirard, B.; Pickett, S. D. *J. Chem. Inf. Comput. Sci.*, **2000**, *40*, 1431.
[92] Gao, H. *J. Chem. Inf. Comput. Sci.*, **2001**, *41*, 402.
[93] van Rhee, A. M.; Stocker, J.; Printzenhoff, D.; Creech, C.; Wagoner, P. K.; Spear, K. L. *J. Comb. Chem.*, **2001**, *3*, 267.
[94] Nicolaou, C. A.; Tamura, S. Y.; Kelley, B. P.; Bassett, S. I.; Nutt, R. F. *J. Chem. Inf. Comput. Sci.*, **2002**, *42*, 1069.
[95] Wilton, D.; Willett, P.; Lawson, K.; Mullier, G. *J. Chem. Inf. Comput. Sci.*, **2003**, *43*, 469.
[96] Charifson, P. S.; Walters, W. P. *J. Comput. Aided Mol. Des*, **2002**, *16*, 311.

[97] Muegge, I.; Heald, S. L.; Brittelli, D. *J. Med. Chem.*, **2001**, *44*, 1841.
[98] Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. *J. Comb. Chem.*, **1999**, *1*, 55.
[99] Bajorath, J. *Nat. Rev. Drug Discov.*, **2002**, *1*, 882.
[100] Molnar, L.; Keseru, G. M. *Bioorg. Med. Chem. Lett.*, **2002**, *12*, 419.
[101] Engels, M. F.; Thielemans, T.; Verbinnen, D.; Tollenaere, J. P.; Verbeeck, R. *J. Chem. Inf. Comput. Sci.*, **2000**, *40*, 241.
[102] Godden, J. W.; Furr, J. R.; Bajorath, J. *J. Chem. Inf. Comput. Sci.*, **2003**, *43*, 182.
[103] Engels, M. F.; Venkatarangan, P. *Curr. Opin. Drug Discov. Devel.*, **2001**, *4*, 275.