Data structures and computational tools for the extraction of SAR information from large compound sets

Mathias Wawer[1], Eugen Lounkine[1], Anne M. Wassermann and Jürgen Bajorath

Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, D-53113 Bonn, Germany

Computational data mining and visualization techniques play a central part in the extraction of structure–activity relationship (SAR) information from compound sets including high-throughput screening data. Standard statistical and classification techniques can be used to organize data sets and evaluate the chemical neighborhood of potent hits; however, such methods are limited in their ability to extract complex SAR patterns from data sets and make them readily accessible to medicinal chemists. Therefore, new approaches and data structures are being developed that explicitly focus on molecular structure and its relationship to biological activity across multiple targets. Here, we review standard techniques for compound data analysis and describe new data structures and computational tools for SAR mining of large compound data sets.

Advances in high-throughput screening (HTS) technology and combinatorial chemistry have led to the fast accumulation of large amounts of activity data for chemical compounds [1]. Extracting SAR information from such data sets and making it available to medicinal chemists are important tasks in hit selection and hit-to-lead projects. To extract SAR information from HTS data or other compound data sets, various computational data mining and visualization approaches have been developed over the years. These methods have in common that computer-readable molecular representations must be used. To make chemical structure accessible to computational approaches, molecules are represented, for example, as vectors of numerical descriptors or molecular fingerprints that account for the presence of predefined or calculated structural features. Given these inherently multi-dimensional representations and the availability of large amounts of activity and other biological data, successful chemical data mining must ultimately find a balance between mathematical complexity reduction and chemical interpretability. The primary goal is to systematically extract interesting data set features and present them in an intuitive way.

Corresponding author. Bajorath, J. (bajorath@bit.uni-bonn.de)

[1] These authors contributed equally to the article.

Four conceptually different approaches to SAR extraction from large data sets can be distinguished: dimensionality reduction, clustering and partitioning, organization and annotation of compound substructures, and relating structural similarity to activity similarity. In the first part of this review, we provide an overview of each of these approaches and give examples of the applied mathematical methods, data structures and computational tools. In the second part, we describe in detail data structures that put strong emphasis on integrating structural and activity similarity with a focus on methods recently developed in our group.

Data preprocessing

Before an in-depth analysis, it is usually required to curate and combine data from distinct sources, which makes preprocessing an important step in SAR information extraction. This is especially true for HTS data, which – owing to the high level of automation – require careful control and analysis to reliably identify hits for further evaluation. A detailed discussion of these issues is outside the scope of this article but has been presented elsewhere [2,3].

For the recurrent tasks of data preprocessing, pipelining programs such as Pipeline Pilot (http://accelrys.com/products/pipeline-pilot/) and KNIME [4] (http://www.knime.org) are useful and versatile tools. Protocols for standard processing tasks can be implemented quickly and existing workflows easily adapted to

1359-6446/06/$ - see front matter doi:10.1016/j.drudis.2010.06.004

new requirements, which has made these tools widely used in chemoinformatics.

Dimensionality reduction

To efficiently mine and recognize patterns in multi-dimensional data, the data must be represented in a human-accessible format. To this end, one possibility is to reduce high-dimensional data representations to a readily interpretable two- or three-dimensional reference space that can be intuitively navigated. These low-dimensional reference spaces can be visualized using standard techniques such as scatter plots in which each data point corresponds to a molecule. Annotation of data points according to activity or other molecular properties enables the identification of regions in reduced chemical space that are predominantly populated by active compounds. Using computer programs such as Spotfire [5], the data can be projected onto multiple plots representing chemical and activity space. Series of compounds that provide SAR information are then identified by comparing the data points in different diagrams; however, compound structures must be separately compared and interpreted to extract meaningful SAR information. Thus, in such situations, chemical experience and intuition are of cardinal importance for data analysis.

Various mathematical methods can be applied to facilitate dimensionality reduction. For more detailed information concerning dimension reduction techniques, interested readers are referred to Refs. [6–8].

Dimension reduction is generally accompanied by some loss of the original information content and feature variance of high-dimensional representations of a data set. For example, principal component analysis (PCA) generates linear combinations of original descriptors, and a small number of these linear combinations, utilized as new variables, often explain a large fraction of the data set variance. The web service ChemGPS uses a predefined principal components space that has been established on the basis of bioactive compounds [9,10]. A related approach uses Bayesian modeling to define a low-dimensional bioactivity space into which compounds can be projected. For this purpose, the so-called 'Bayesian affinity fingerprints' are generated by predicting targets of compounds based on their structure and properties. Principal components of the prediction scores ultimately constitute a low-dimensional bioactivity space [11].

Multi-dimensional scaling aims to retain high-dimensional pairwise similarity relationships in low-dimensional space representations. This has the advantage that close similarity relationships are often better represented than by PCA; however, a computed mapping cannot be transferred to another data set because it derives coordinates based on relative compound distances for a given data set.

Kohonen networks or self-organizing maps represent a complex methodology for dimension reduction [12]. Network training generates a two-dimensional map containing regions populated with structurally related compounds. Activity information can be projected by coloring individual cells on this map. Feed-forward neural networks are supervised learning algorithms that can be trained to reduce dimensionality in a favorable way (e.g. by distinguishing active from inactive compounds).

All of these dimension reduction methods have in common that molecules corresponding to interesting data points in low-dimensional representations need to be compared manually to extract chemically interpretable SAR information.

Clustering and partitioning

The general aim of clustering and partitioning methods in compound data analysis is to group structurally similar compounds together. Often, clusters contain analogs representing the same chemotype. The biological activities of analogs or compounds representing similar chemotypes can then be compared (Fig. 1b). These data analysis techniques typically operate in high-dimensional chemical reference spaces; however, partitioning can be elegantly combined with dimension reduction [13]. In general, clustered structures must also be visually inspected and compared to recognize SAR patterns. Agglomerative hierarchical clustering uses pairwise compound similarity in a high-dimensional chemical space to organize compounds based on common structural elements. Most often, the clustering is visualized as a dendrogram where each node corresponds to a particular cluster at a defined similarity level. The program Molecular Property eXplorer [14] uses this type of clustering in combination with the so-called 'tree maps' to visualize the global structure of the data set. In tree maps, hierarchical clustering is represented as a rectangular map that is sub-divided according to the grouping of compounds. An advantage of tree maps over dendrograms is that distances between compounds in different clusters are also visualized [14]. Each sub-rectangle represents a cluster, and the rectangles are colored according to bioactivity. To incorporate more than one property, Molecular Property eXplorer uses heatmaps that combine clustering of compounds with clustering of properties. Each cell is colored according to the property of the given compounds. This enables straightforward visual identification of compound clusters that share a common activity or are heterogeneous with respect to a certain property.

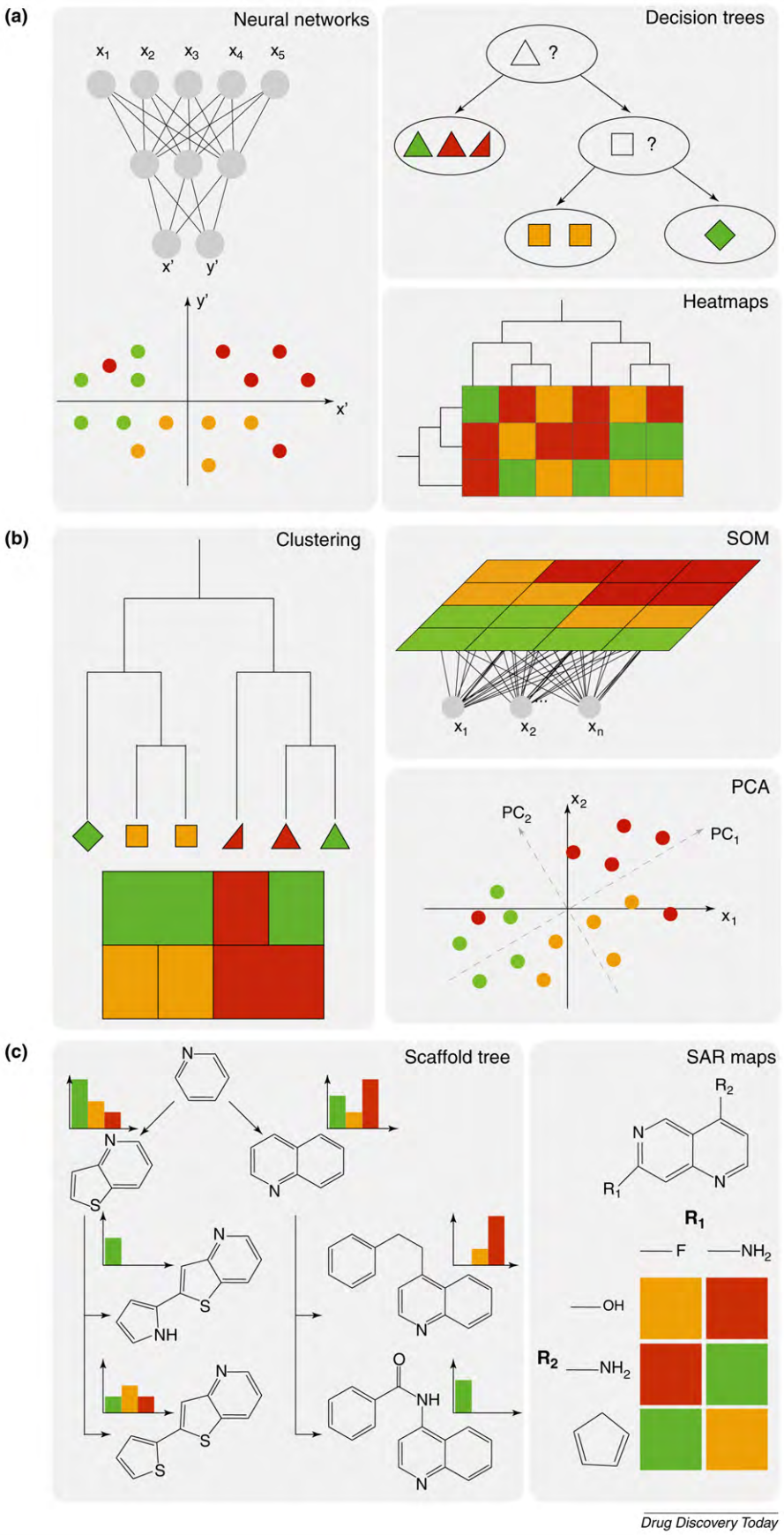
Whereas clustering requires pairwise comparison of all compounds in high-dimensional chemical space, partitioning methods separate the data set based on individual property ranges or the presence of selected structural features. For example, in decision trees the data set is first split according to a single property so that the separation into actives and inactives is best at that stage [15,16]. The process is then repeated for each compound subset. Rules concerning structural features and chemical properties that are markers of activity can then be derived from decision trees.

A grouping of compounds by their chemical features and topology can also be obtained using reduced graphs [17,18]. These graph-based descriptions can be evolved to capture characteristic features of active molecules that yield interpretable SAR patterns.

We can only give an overview of clustering and partitioning methods in this review. For further information, useful publications about standard data mining techniques can be consulted [8,19,20].

Organization and annotation of substructures

Substructure-centric approaches aim to reveal structural rules that govern bioactivity to guide compound design. Therefore, substructures are either predefined or generated in a systematic manner from a compound data set. Each substructure is then annotated with the activity of the compounds it represents (Fig. 1c).



Drug Discovery Today

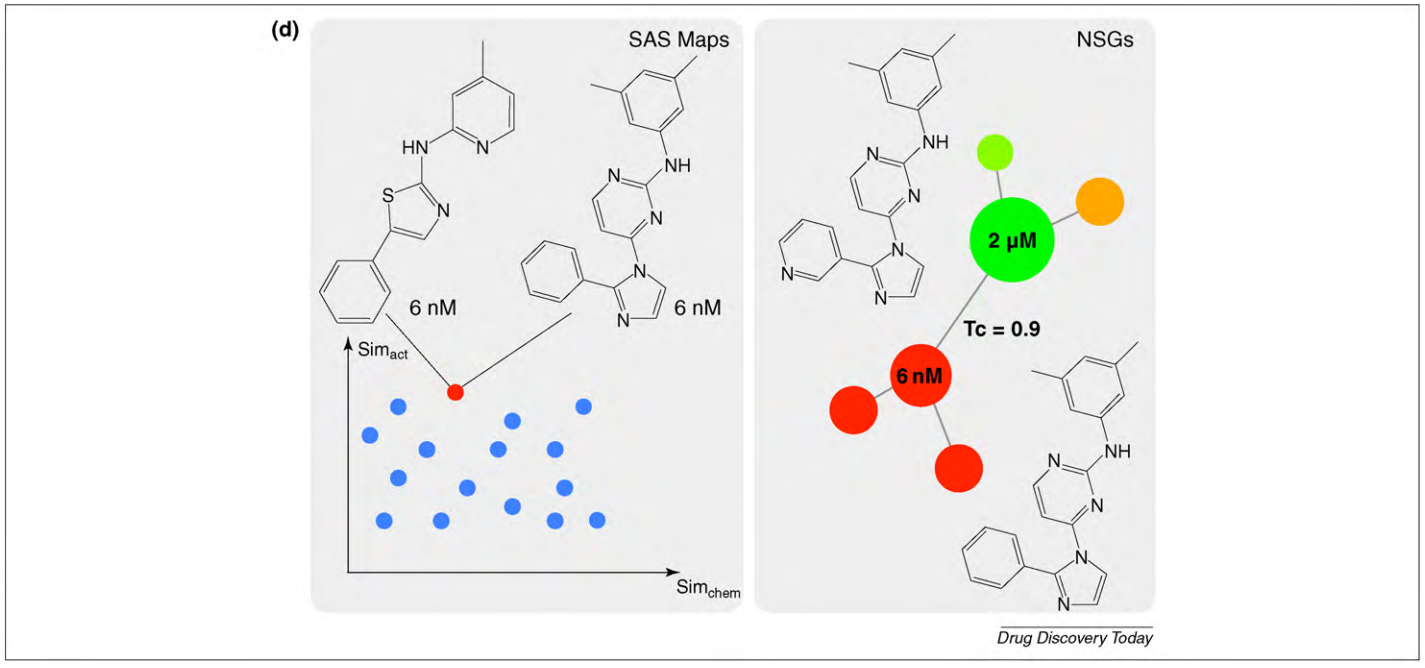


Fig. 1. (*Continued*).

LeadScope [21] uses a library of predefined substructures that often occur in drug-like compounds. The program permits the selection of substructures that occur in compounds with defined activity and property ranges. This makes it possible to visualize SAR information in terms of chemically intuitive structural features.

Another program, Enhanced SAR Maps, uses heatmap representations to visualize the distribution of activity and other properties such as toxicity or bioavailability for analog series representing a particular chemotype [22,23]. These heatmap views enable the visual identification of activity patterns that are associated with different chemical substitutions. This approach is closely related to SAR tables that are commonly used by medicinal chemists to summarize SARs, but the analysis is limited to one common core structure at a time.

Scaffold Hunter [24] uses a data structure called scaffold tree [25] in which molecules are systematically and iteratively decomposed into substructures. Each substructure is annotated with the bioactivities of compounds it was generated from. The organization of compound data sets into molecular building blocks of systematically varying size provides a structural hierarchy for activity annotation. The scaffold tree enables the identification of scaffolds that are characteristic of active compounds. It has been shown that utilizing activity-prevalent substructures as scaffolds for newly synthesized molecules can yield novel active compounds [24].

The structure–activity report that is part of the Molecular Operating Environment (http://www.chemcomp.com) combines a scaffold-tree-based grouping of compounds with SAR table features similar to SAR Maps, as well as other graphical representations. It enables the automatic identification of scaffolds and provides a detailed graphical view of the distribution of one or multiple properties among a set of compounds [26].

Structural *vs.* activity similarity

In contrast to the methods described thus far, the approaches discussed in the following paragraphs are designed to integrate structural similarity and activity similarity and provide a consistent framework for the extraction and interpretation of SAR information (Fig. 1d). Two basic SAR categories can be distinguished: continuous and discontinuous SARs. Discontinuity is introduced when structurally similar molecules have very different potency; pairs or groups of such compounds form 'activity cliffs'. By contrast, increasingly dissimilar molecules with comparable potency form a continuous SAR relationship. Importantly, these SAR categories are not mutually exclusive and SARs combining continuous

FIGURE 1

Method for extracting and organizing SAR information. Four conceptually different approaches to SAR information extraction from compound and HTS data are shown. **(a)** Dimension reduction. Neural networks can be trained to reduce dimensionality in a well-defined manner. Using self-organizing maps (SOMs), compounds are projected onto a two-dimensional map (and colored according to potency or other properties). Principal component analysis (PCA) is used to reduce the dimensionality of descriptor spaces by constructing composite descriptors from the original ones. **(b)** Clustering and partitioning. Hierarchical clustering separates compounds into groups of similar structures. The cluster organization can be visualized as a dendrogram or tree map. Heatmaps combine clustering of properties and compounds. Decision trees derive rules to enrich active compounds in terminal nodes. **(c)** Organization and annotation of substructures. The scaffold tree represents an iterative hierarchical fragmentation and organization scheme of compounds. Individual scaffolds are annotated with the activity of molecules they represent. SAR maps report activity in relation to combinations of functional groups attached to a common core structure. **(d)** Structural *vs.* activity similarity. SAS maps report the structural and activity similarity of compound pairs in a scatter plot. This enables the identification of activity cliff markers. Network-like similarity graphs (NSGs) visualize both structural and activity similarity for an entire data set.

TABLE 1

SAR information extraction methods.

Approach	Data structures and methods	Software tools
Dimensionality reduction	Principal component analysis Self-organizing mapsNeural networks	Spotfire
Clustering and partitioning	Dendrograms Tree mapsHeatmaps Decision trees	Molecular Property eXplorer
Substructure annotation	Scaffold treeSAR Maps	LeadScope Scaffold Hunter Third Dimension Explorer (3DX)
Similarity comparison	Spiral viewsSALI, SARI Network-like similarity graphs Combinatorial analog graphs	Spiral View Saranea

and discontinuous components are classified as heterogeneous. The designation of these categories originates from the concept of an 'activity landscape' that describes the potency distribution of a compound set with reference to the distance of its members in chemical space. When chemical space is projected onto a two-dimensional plane where inter-compound distances reflect structural similarity, potency can be added as a third dimension to create a surface with smooth (continuous) and cliffy (discontinuous) regions, akin to a real geographical landscape. The distinction of these SAR categories promotes a more systematic approach to the description of SARs and their relevance for different applications. Discontinuous behavior, for example, is often exploited in lead optimization when high gains in potency are achieved by small structural modifications. By contrast, continuous SAR character is thought to be a prerequisite for the applicability of virtual screening, lead hopping or QSAR techniques [27].

Fundamental to all SAR analysis methods that employ this SAR categorization are pairwise comparisons of molecules. Structure–activity similarity (SAS) maps [27,28] depict all possible compound pairs of a data set in a scatter plot that reflects the relationship between structure and potency similarity. These two-dimensional representations permit the identification of pairs or groups of compounds that exhibit either continuous or discontinuous behavior and provide an overview of their distribution over an entire data set. More detailed analysis can be difficult, however, because the basic units of representation in SAS maps are compound pairs (i.e. each data point corresponds to a pair). No immediate visual access is possible to individual molecules or pair relationships (for example, the set of all pairs formed by a particular compound).

The latter aspect, the exploration of the structural neighborhood of a given compound, is the focus of a highly interactive method termed spiral view (SV) [29] that was originally conceived for the analysis of activity cliffs. Around a manually selected compound of interest, chemical neighbors are positioned according to their similarity to the central compound. Links between compound depictions indicate to what extent these compounds differ in activity or other properties. The user can also select a neighbor compound of interest to generate a new SV around it. Thus, property changes related to structural changes of varying degrees can be explored; however, this methodology is not capable of representing a large data set.

Two numerical SAR analysis functions have been developed for the quantitative description of SAR continuity and discontinuity that combine structural and activity similarity in one metric. Analogous to SVs and SAS maps, these functions focus on different characteristics of an activity landscape. The structure–activity landscape index (SALI) [30] quantifies the extent of discontinuity for individual compound pairs. Contiguous activity cliffs can be visualized in a graph in which molecules (represented as nodes) are connected if they have similar structures but significantly different potency. Structural transitions between individual molecules are also captured by the concept of matched molecular pairs [31] and similar approaches [32] that aim to identify substructural differences between similar molecules. Such approaches can thus also be used to identify transitions that are associated with activity cliffs.

By contrast, the SAR index (SARI) function [33] characterizes the SAR phenotype of an entire data set. SARI values range from 0 (purely discontinuous) to 1 (purely continuous), and intermediate values indicate varying degrees of SAR heterogeneity.

Network-like similarity graphs (NSGs) [34] combine the representation of pairwise molecular relationships in a graph (as in SVs or the SALI graphs) with a holistic view of a data set (as in SAS maps or SARI analysis). In the following paragraphs, we will focus on the evolution of data structures for SAR analysis that we have derived from NSGs to highlight several features that might also be useful for the design of other SAR data visualization techniques. Table 1 lists methods and software tools representing the four principal approaches to SAR-relevant compound data mining.

Evolution of SAR data structures

A primary aim of visualization in data mining is to provide an unbiased representation of the underlying data while reducing the complexity of the data representation. This can be achieved by graphically emphasizing selected characteristics of a data set or by showing only parts of the data. The decision about feature significance ultimately depends on the specific goals of the analysis and determines the quality of the resulting visualization. The primary purpose of SAR analysis is to elucidate how structural features of small molecules are related to their biological activity. Accordingly, the SAR analysis techniques described in the following paragraphs are based on systematic comparisons of compound structure and potency. Crucial aspects for the design of graphical SAR analysis

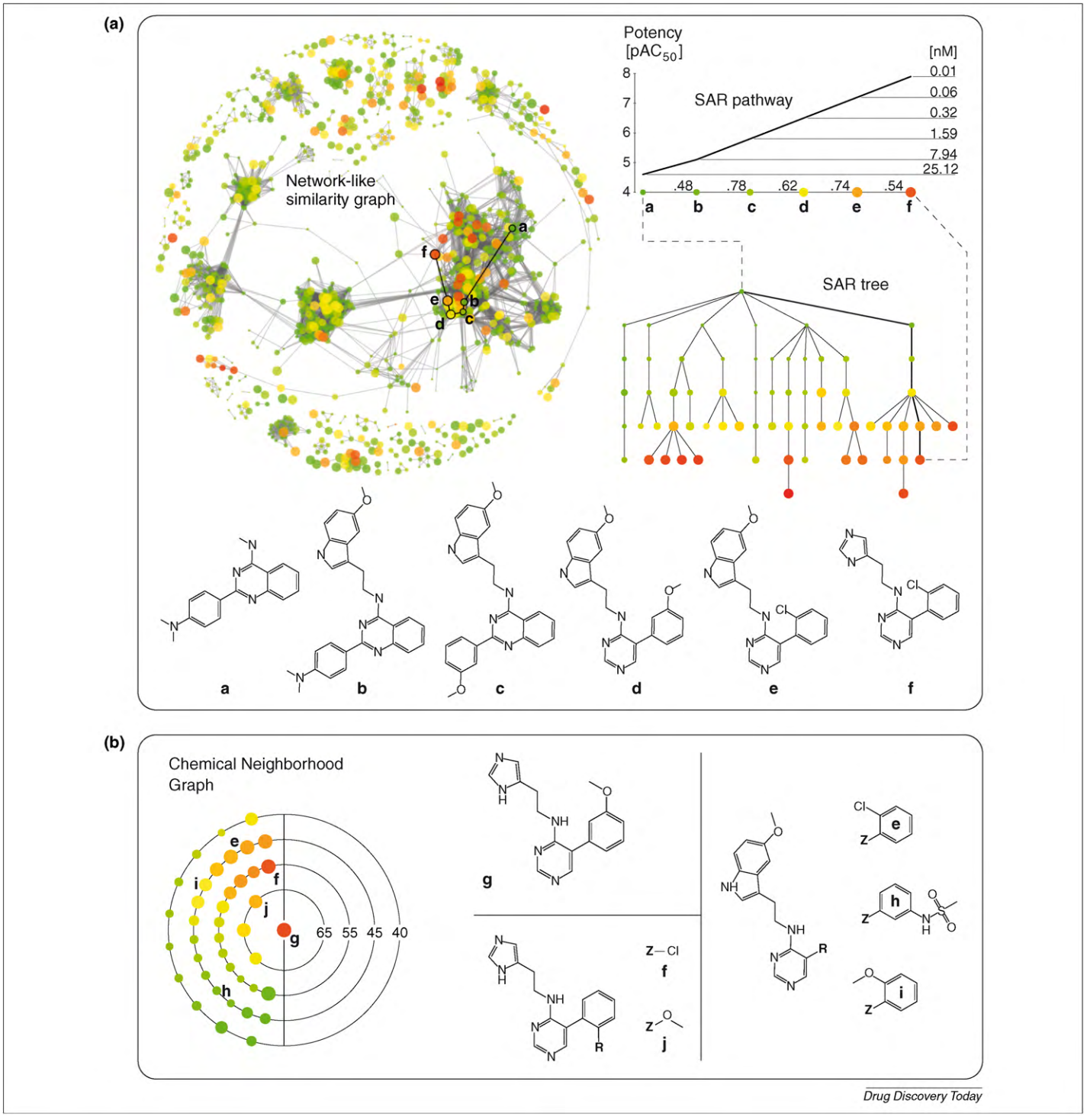


FIGURE 2

Graph-based methods for SAR information extraction. Examples of an NSG and NSG-derived data structures are shown for a data set of cytochrome P450 isoform 2C19 inhibitors taken from PubChem. **(a)** On the left, an NSG is shown that represents the complete data set. A pathway of compounds is highlighted and shown in detail in the top right corner. Below the pathway plot, a SAR tree is presented and the selected pathway is traced using thick lines. The compounds in the lower part of the figure correspond to the nodes forming the pathway. Their positions in the three data structures are indicated using a letter code. **(b)** An exemplary Chemical Neighborhood Graph (CNG) and selected compounds are shown. Compound labeling is consistent for parts (a) and (b) so that compounds that occur in several data structures are identified.

tools include which molecules to select and how to compare them. The data structures discussed in the next section have in common that they are based on pairwise molecular comparisons and similarity-based graph or network representations.

Network-like similarity graphs

Network-like similarity graphs, mentioned above, represent a compound data set by showing all molecules and their similarity relationships (Fig. 2a). NSGs are formal graphs in which nodes

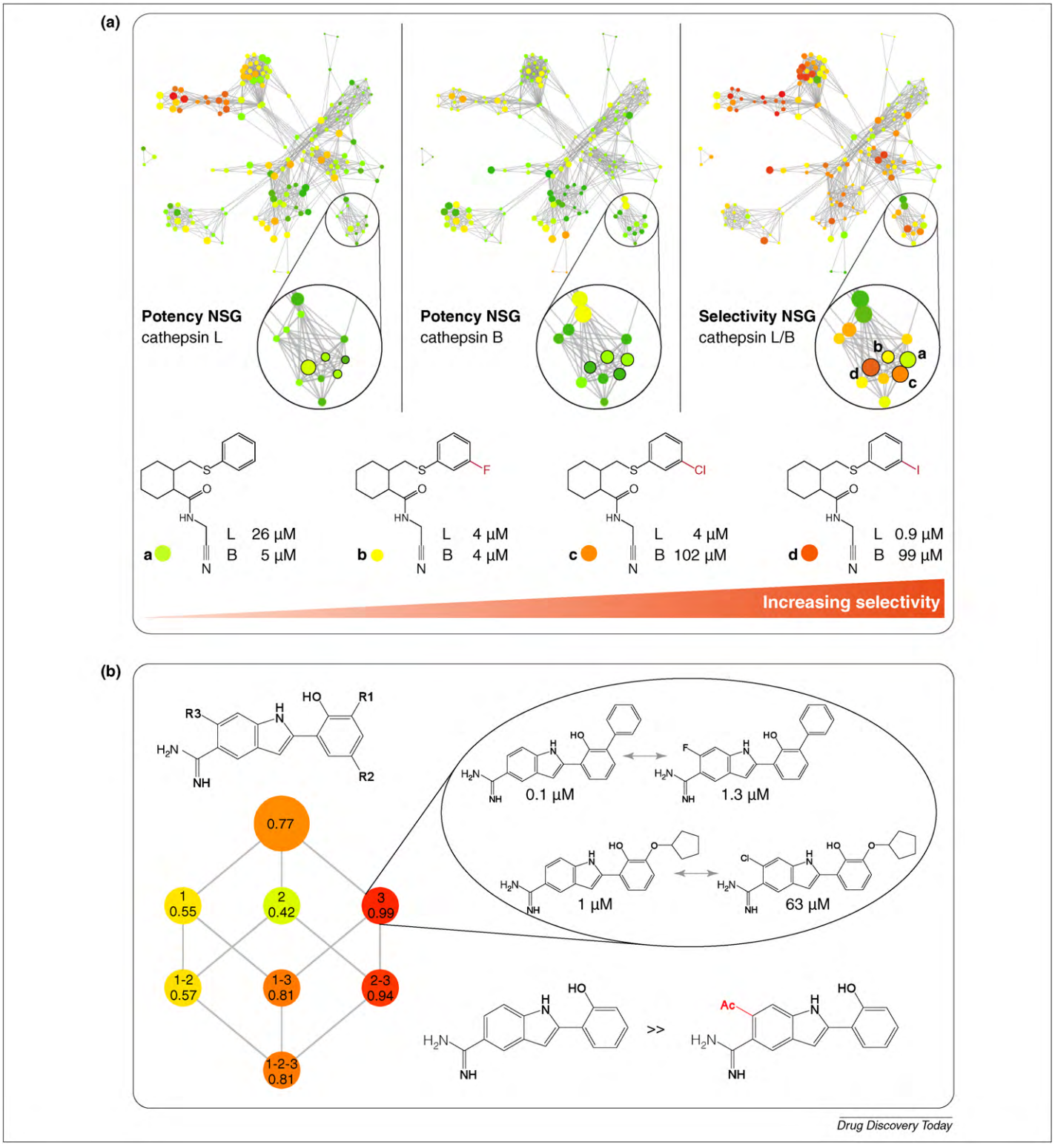


FIGURE 3

Multi-target SAR analysis. **(a)** NSGs. Three NSGs with identical layout are shown that visualize the same set of cathepsin inhibitors. On the left and in the center, nodes are color coded by compound potency for two different targets: cathepsin L (left) and cathepsin B (center). In these potency-based NSGs, node size reflects compound discontinuity scores calculated for potency values. By contrast, the right graph captures selectivity of inhibitors for cathepsin L over cathepsin B. In this selectivity-based NSG, red nodes correspond to compounds that are selective for cathepsin L and green nodes to those selective for cathepsin B. Yellow nodes correspond to non-selective inhibitors. In this case, node size reflects compound discontinuity scores for selectivity values (i.e. logarithmic potency differences). Four compounds with different selectivity behavior have been selected from the highlighted sections of the graphs and are shown at the bottom of the figure. Inhibitor selectivity for cathepsin L increases from the left to right and halogen atoms of increasing size and electronegativity become apparent selectivity determinants. **(b)** Combinatorial analog graphs. A CAG representation is shown for a factor Xa inhibitor analog series with three substitution sites. Nodes are colored according to compound subset discontinuity scores from green (low SAR discontinuity) to red (high discontinuity). Site 3 is identified as a 'SAR hotspot' where different substituents have dramatic influence on compound potency. The analysis of pharmacophore feature replacements for compound pairs that differ

correspond to molecules. Pairwise similarity relationships are represented by edges that connect individual nodes. Only molecule pairs that exceed a predefined similarity threshold are connected by an edge. To visualize the potency distribution, nodes are color coded by potency, applying a continuous spectrum from green (lowest potency in the data set) to red (highest potency). Accordingly, similarity relationships determine the structure of the graph, whereas potency is used as an annotation of nodes. In contrast to the methods belonging to the first three categories, as discussed above, the NSG data structure has been specifically designed to enable both global and local data set analysis and compound comparisons. A graphical layout algorithm is applied to arrange sets of similar molecules as separated clusters, which enables the analysis of the interplay between global and local SAR features (Fig. 2a). NSGs contain multiple annotations representing different layers of information. To support the interpretation of SARs, the diagrams are annotated with three numerical scores that reflect SAR characteristics of the entire data set, compound clusters (subsets) and individual molecules. The SARI function is used to categorize the global SAR type formed by all compounds in a data set as continuous, discontinuous or heterogeneous. SARI cluster scores are calculated to quantify the level of discontinuity in different subsets of similar molecules; high cluster scores indicate the presence of multiple activity cliffs. The compound discontinuity score is represented by node scaling and reflects the potency deviation of a compound from its structural neighbors. Thus, it identifies molecules that introduce SAR discontinuity and activity cliffs. In NSGs, combinations of large red and green nodes connected by an edge are activity cliff markers that can be easily identified. NSGs enable the selection of compounds that determine both local and global SAR features or compounds that are prime candidates for chemical exploration or optimization efforts because they bridge between continuous and discontinuous local SAR environments. However, NSGs are generally too complex for the extraction of detailed SAR rules or derivation of specific hypotheses for compound design. For these purposes, three data structures that mine the information contained in NSGs in different ways have been developed. All of them are subgraphs of these networks (i.e. they contain only a subset of the nodes and edges in an NSG but use the same elements to convey information – nodes, edges, node color and node size). Thus, NSGs and their derivatives are designed for consistent data representation and interpretation.

SAR pathways and trees

The first in this series of data structures are the so-called 'SAR pathways' [35]. They were conceptualized to capture potency effects accompanying stepwise structural changes and consist of sequences of pairwise similar compounds (nodes). Consequently, they correspond to a path through contiguous edges in an NSG (Fig. 2a). Evaluating all possible pathways in an NSG is not feasible and thus a model was defined for preselection of favorable pathways. Given two nodes within an NSG, only the best pathway connecting them according to the underlying model is evaluated. This model identifies regions where structural similarity is reflected by potency similarity (i.e. where gradual changes in potency along a series of similar molecules occur). Essentially, this pathway model reflects SAR continuity, and pathways best fitting these criteria are automatically extracted. From preferred pathways, interesting compound series can quickly be identified (at the expense of a comprehensive view of the underlying SAR, however). Importantly, activity cliffs are by design not covered by the pathway model; however, all pathways leading to a particular activity cliff can be selected and compared, which connects regions of local SAR continuity and discontinuity. Furthermore, pathways might include scaffold hops because they are formed by pairwise similar compounds, which makes structural transition along a pathway possible.

SAR pathways that originate or end at the same compound can be arranged as branches of the so-called 'SAR trees' (Fig. 2a) that share the SAR model dependency of individual pathways. The common start or end point forms the tree root and identical sections of different paths are fused into the same branch. SAR trees provide a structural context for individual pathways and compare alternative routes to compound modification.

Chemical neighborhood graphs

The third data structure derived from NSGs departs from analyzing consecutive compound modifications and promotes a compound-centric view of SARs. Chemical neighborhood graphs (CNG) [36] visualize the similarity and potency distribution in the structural neighborhood of a given compound (Fig. 2b). The neighborhood of a reference compound is defined as the set of compounds that exceed a predefined similarity threshold to the reference compound and populate a similarity radius around it. Although developed independently, CNGs are reminiscent of SVs. Different from the other NSG-derived data structures, CNGs do not contain edges and are thus a set of nodes rather than a graph, but the meaning of node color and size remains unchanged. Although similarity relationships are not explicitly represented by edges, they are fundamental to the construction of CNGs. In the graphical representation, all nodes of a neighborhood are arranged on concentric circles around a central node corresponding to the reference molecule. Each circle represents a unique range of similarity values and the radii of circles reflect the order of value ranges. The structurally most similar molecules are placed closest to the center on the smallest circle, whereas compounds close to the similarity threshold are located on the outer (and largest) circle.

CNGs provide a detailed overview of the potency distribution in a compound's structural neighborhood and are particularly easy to interpret, given their similarity-based design. In contrast to SAR pathways and trees, no predefined SAR model is used for their construction. CNGs are systematically computed for each compound in a data set and ranked by simple parameters to identify the most information-rich and interpretable neighborhoods. CNGs present overlapping compound neighborhoods, which often helps to view local SAR information from

at site three reveals that the introduction of acyclic substituents (Ac) significantly decreases potency. Acyclic substituents would hence obtain low priority in the preference order for this substitution site.

different perspectives (relative to different reference compounds) and substantially aids in the interpretation of complex SAR features.

A common feature of the graphical analysis methods derived from NSGs is that they are amenable to multi-parametric data visualization. Therefore, we present in the following paragraphs examples of advanced applications addressing the analysis of multi-target SARs (mtSARs).

Selectivity NSGs

There is increasing notion of polypharmacological drug behavior [37] and accumulating evidence that target selectivity often results from differences in compound potency against multiple targets, rather than from specific single-target interactions [38]. Accordingly, the study of mtSARs is expected to become increasingly relevant in the future because it provides a basis for the identification of molecular selectivity determinants. As a first attempt to explore compound selectivity in a systematic and quantitative manner, the SARI scoring scheme and NSGs were adapted to analyze compound data sets annotated with potency information for two or more targets [39]. Initially, potency-centric NSGs are generated for the individual targets. Furthermore, to enable dual-target SAR or 'structure–selectivity relationship' (SSR) analysis, selectivity values of active compounds are calculated as the difference between their logarithmic potency values for two targets, then global and local SARI scores are calculated on the basis of selectivity values and a selectivity-centric NSG is generated. In selectivity NSGs, the color code of nodes reflects selectivity values. In addition, high compound discontinuity scores (reflected by large node sizes) identify molecules that display notable differences in target selectivity compared to their structural neighbors and hence form 'selectivity cliffs'. Because the topology of NSGs is only determined by similarity relationships between the compounds in the data set, it is conserved in the selectivity- and potency-centric graph representations (Fig. 3a), which makes it possible to directly compare corresponding compounds. Hence, by side-by-side comparison of corresponding regions in the potency and selectivity NSGs, key compounds that influence SSR and single-target SARs in similar or different ways can easily be identified. This makes it possible to prioritize compounds based on their SAR and SSR features (Fig. 3a).

Saranea

To make NSG-based data mining techniques publicly available to the scientific community, we have recently released the Saranea program [40] (SAR/*Araneae*; i.e. the scientific order of spiders, reminding us of spider webs and networks). Saranea provides a graphical user interface to NSGs and NSG-based data structures and accepts customized molecular fingerprint representations and potency data as input. The central feature of Saranea is the simultaneous interactive exploration of multiple NSGs. Compound selection in one (potency or selectivity) NSG representation can be automatically synchronized with all other NSG views, making it possible to compare multiple SARs and SSRs for different targets. The program also integrates user-defined descriptors and depicts molecules, providing immediate interactive access to the molecular structures represented by nodes in the graphs. SAR and SSR trees, pathways, and chemical neighborhood graphs are integrated for interactive tool editing of pathways using information provided by CNGs.

The program is freely available, together with its source code, providing opportunities for others to use, extend and, we hope, further develop SAR visualization techniques on the basis of NSGs and related data structures [40].

Combinatorial analog graphs

Although the inspection of the environments of key compounds in selectivity NSGs provides a sound basis for the identification of selectivity determinants, for lead optimization purposes, a data structure specifically designed to focus on analog series is also desirable. To this end, combinatorial analog graphs (CAGs) [41] have been developed that analyze SARs at the level of individual substitution sites and their combinations. CAG analysis was recently extended to the study of mtSARs [42]. Following this approach, the maximum common subgraph of an analog series is determined and used as a core structure for R-group decomposition of individual analogs such that substituents are assigned to consistently numbered substitution sites. SARI discontinuity scores are then calculated for subsets of compounds that only differ at a defined substitution site or combinations of up to three sites. These subsets correspond to nodes in CAG representations, which sets this data structure conceptually apart from NSG-based structures in which nodes correspond to individual compounds (Fig. 3b). Accordingly, edges in CAGs connect subsets and indicate that compounds in these subsets share modifications at the same substitution sites. Furthermore, nodes are color coded according to discontinuity scores and not according to potency. High discontinuity scores highlight SAR hotspots (i.e. substitution sites where the modification of functional groups leads to significant differences in potency and the introduction of SAR discontinuity) (Fig. 3b). Compound subsets of key sites and SAR hotspots are selected for further analysis. To gain a better understanding of structural features that are responsible for significant changes in potency, substituents are encoded as predefined pharmacophore features and compound pairs are grouped according to the pharmacophore feature replacements required to convert one compound into the other. All potency changes associated with specific pharmacophore feature exchanges are recorded and enable the derivation of 'preference orders' for given substitution sites. For the same analog series, this analysis can be carried out for multiple targets and preference orders can be compared across these targets, leading to the identification of differences in preferred substitutions at given sites or site combinations. These differences identify pharmacophore features that can serve as selectivity determinants and can be exploited to increase compound selectivity for one target over one or more others. Hence, on the basis of comparative CAG analysis, simple and intuitive rules can often be formulated to guide the design of target-selective compounds.

Concluding remarks

The extraction of SAR information from large compound data sets has a crucial role during the early stages of drug discovery. Herein, we have reviewed conventional and newly developed approaches to search for and rationalize SAR information and aid in the selection of active compounds for further chemical exploration. New approaches to large-scale SAR analysis employ data mining

methods and emphasize two aspects: systematic analysis and intuitive graphical representation of SAR features. The methodological framework presented herein provides a basis for the integration of data mining approaches and graphical analysis techniques to complement chemical knowledge and experience and formulate SAR hypothesis for experimental design. It is anticipated that recently introduced approaches focusing on systematic comparison of chemical and activity similarity with emphasis on graphical accessibility will catalyze further development in the SAR analysis field that has evolved at the interface between computational and medicinal chemistry.

Acknowledgement

The authors thank Lisa Peltason for help with Fig. 3.

References

1 Mayr, L.M. and Bojanic, D. (2009) Novel trends in high-throughput screening. *Curr. Opin. Pharmacol.* 9, 580–588
2 Malo, N. *et al.* (2006) Statistical practice in high-throughput screening data analysis. *Nat. Biotechnol.* 24, 167–175
3 Harper, G. and Pickett, S.D. (2006) Methods for mining HTS data. *Drug Discov. Today* 11, 694–699
4 Berthold, M.R. *et al.* (2008) Knime: the konstanz information miner. In *Data Analysis, Machine Learning and Applications* (Preisach, C. *et al.* eds), pp. 319–326, Springer
5 Ahlberg, C. (1999) Visual exploration of HTS databases: bridging the gap between chemistry and biology. *Drug Discov. Today* 4, 370–376
6 Ivanenkov, Y.A. *et al.* (2009) Computational mapping tools for drug discovery. *Drug Discov. Today* 14, 767–775
7 Agrafiotis, D. and Lobanov, V.S. (2000) Nonlinear mapping networks. *J. Chem. Inf. Comput. Sci.* 40, 1356–1362
8 Hair, J.F. *et al.* eds (1998) *Multivariate Data Analysis*, Prentice-Hall
9 Oprea, T.I. and Gottfries, J. (2001) Chemography: the art of navigating in chemical space. *J. Comb. Chem.* 3, 157–166
10 Larsson, J. *et al.* (2005) Expanding the ChemGPS chemical space with natural products. *J. Nat. Prod.* 68, 985–991
11 Bender, A. *et al.* (2006) 'Bayes affinity fingerprints' improve retrieval rates in virtual screening and define orthogonal bioactivity space: when are multitarget drugs a feasible concept? *J Chem. Inf. Model.* 46, 2445–2456
12 Yan, A. (2006) Application of self-organizing maps in compounds pattern recognition and combinatorial library design. *Comb. Chem. High Throughput Screen* 9, 473–480
13 Bajorath, J. (2002) Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discov.* 1, 882–894
14 Kibbey, C. and Calvet, A. (2005) Molecular property eXplorer: a novel approach to visualizing SAR using tree-maps and heatmaps. *J. Chem. Inf. Model.* 45, 523–532
15 Cross, K.P. *et al.* (2003) Finding discriminating structural features by reassembling common building blocks. *J. Med. Chem.* 46, 4770–4775
16 Miller, D.W. (2003) A chemical class-based approach to predictive model generation. *J. Chem. Inf. Comput. Sci.* 43, 568–578
17 Birchall, K. *et al.* (2008) Evolving interpretable structure–activity relationships. 1. Reduced graph queries. *J. Chem. Inf. Model.* 48, 1543–1557
18 Birchall, K. *et al.* (2008) Evolving interpretable structure–activity relationship models. 2. Using multiobjective optimization to derive multiple models. *J. Chem. Inf. Model.* 48, 1558–1570
19 Jain, A.K. *et al.* (1999) Data clustering: a review. *ACM Comput. Surv.* 31, 264–323
20 Tan, P. *et al.* (2005) Clustering analysis: basic concepts and algorithms. In *Introduction to Data Mining* (Tan, P. *et al.* eds), pp. 487–568, Addison-Wesley
21 Richon, A. (2000) LeadScope: data visualization for large volumes of chemical and biological screening data. *J. Mol. Graph. Model.* 18, 76–79
22 Agrafiotis, D.K. *et al.* (2007) SAR maps: a new SAR visualization technique for medicinal chemists. *J. Med. Chem.* 50, 5926–5937
23 Kolpak, J. *et al.* (2009) Enhanced SAR maps: expanding the data rendering capabilities of a popular medicinal chemistry tool. *J. Chem. Inf. Model.* 49, 2221–2230
24 Renner, S. *et al.* (2009) Bioactivity-guided mapping and navigation of chemical space. *Nat. Chem. Biol.* 5, 585–592
25 Schuffenhauer, A. *et al.* (2007) The scaffold tree–visualization of the scaffold universe by hierarchical scaffold classification. *J. Chem. Inf. Model.* 47, 47–58
26 Clark, A.M. and Labute, P. (2009) Detection and assignment of common scaffolds in project databases of lead molecules. *J. Med. Chem.* 52, 469–483
27 Bajorath, J. *et al.* (2009) Navigating structure–activity landscapes. *Drug Discov. Today* 14, 698–705
28 Shanmugasundaram, V. and Maggiora, G.M. (2001) Characterizing property and activity landscapes using an information-theoretic approach. *Abstract no. 77, 222nd American Chemical Society National Meeting* Division of Chemical Information
29 Smellie, A. (2007) General purpose interactive physico-chemical property exploration. *J. Chem. Inf. Model.* 47, 1182–1187
30 Guha, R. and Van Drie, J.H. (2008) Structure–activity landscape index: identifying and quantifying activity cliffs. *J. Chem. Inf. Model.* 48, 646–658
31 Leach, A.G. *et al.* (2006) Matched molecular pairs as a guide in the optimization of pharmaceutical properties; a study of aqueous solubility, plasma protein binding and oral exposure. *J. Med. Chem.* 49, 6672–6682
32 Sheridan, R.P. *et al.* (2006) Molecular transformations as a way of finding and exploiting consistent local QSAR. *J. Chem. Inf. Model.* 46, 180–192
33 Peltason, L. and Bajorath, J. (2007) SAR index: quantifying the nature of structure–activity relationships. *J. Med. Chem.* 50, 5571–5578
34 Wawer, M. *et al.* (2008) Structure–activity relationship anatomy by network-like similarity graphs and local structure–activity relationship indices. *J. Med. Chem.* 51, 6075–6084
35 Wawer, M. and Bajorath, J. (2009) Systematic extraction of structure–activity relationship information from biological screening data. *ChemMedChem* 4, 1431–1438
36 Wawer, M. *et al.* (2010) Computational characterization of SAR microenvironments in high-throughput screening data. *Intl. J. High Throughput Screen* 1, 15–27
37 Paolini, G.V. *et al.* (2006) Global mapping of pharmacological space. *Nat. Biotechnol.* 24, 805–815
38 Hopkins, A.L. (2008) Network pharmacology: the next paradigm in drug discovery. *Nat. Chem. Biol.* 4, 682–690
39 Peltason, L. *et al.* (2009) From structure–activity to structure–selectivity relationships: quantitative assessment, selectivity cliffs, and key compounds. *ChemMedChem* 4, 1864–1873
40 Lounkine, E. *et al.* (2010) SARANEA: a freely available program to mine structure–activity and structure–selectivity relationship information in compound data sets. *J. Chem. Inf. Model.* 50, 68–78
41 Peltason, L. *et al.* (2009) Exploration of structure–activity relationship determinants in analogue series. *J. Med. Chem.* 52, 3212–3224
42 Wassermann, A.M. *et al.* (2010) Computational analysis of multi-target structure–activity relationships to derive preference orders for chemical modifications towards target selectivity. *ChemMedChem* 5, 847–858