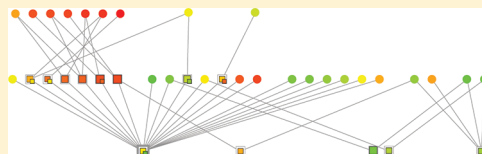


Directed R-Group Combination Graph: A Methodology To Uncover Structure–Activity Relationship Patterns in a Series of Analogues

Anne Mai Wassermann and Jürgen Bajorath*

Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstrasse 2, D-53113 Bonn, Germany

ABSTRACT: A graphical method is introduced to study details of structure–activity relationships (SARs) in analogue series that further extends conventional analysis of analogues using R-group tables or related approaches and that provides additional and more differentiated SAR information. The newly designed graph structure represents entire series of analogues in a consistent manner, regardless of their size and complexity of substitution patterns. The approach is specifically tailored toward a systematic exploration and intuitive interpretation of SAR features involving different R-groups and their combinations. Analogues and their potency information are systematically organized on the basis of R-group combinations that are present in a series. This organization scheme results in graph components that represent well-defined SAR patterns. Analysis of these patterns provides an immediate access to critical substitution sites and R-group combinations, favorable and unfavorable R-groups, or nonadditive potency effects of multisite substitutions. Furthermore, the data structure makes it possible to design new analogues by combining favorable R-group combinations derived from different compounds.



■ INTRODUCTION

Understanding how structural modifications affect the biological activity of small molecules is of central importance in medicinal chemistry. In addition to other approaches to study structure–activity relationships (SARs), computational visualization methods have been introduced that help to extract SAR information from compound data sets.¹ Different methods have been developed to analyze large and diverse compound data sets including high-throughput screening data.^{2,3} The extraction of SAR information from large compound sets represents one of two major tasks in SAR analysis. The other perhaps even more frequently pursued task is compound optimization during later stages of medicinal chemistry efforts. In this case, the focus shifts from larger data sets to individual compound series where SAR exploration primarily aims at the analysis and design of analogues of active compounds with further improved properties. This changes the requirements of SAR exploration and of computational methods employed to aid in this process.¹

The conventional and still most widely used data structure for the analysis of analogue series are R-group tables that contain the core structure common to a series of analogues and rows displaying the substituents of individual compounds and the associated potency measurements. User-friendly extensions of R-group tables have been introduced such as SAR maps⁴ that arrange analogues in rectangular matrices of cells where each cell represents a unique combination of R-groups at two substitution sites. Cells are then color-coded according to a specific molecular property, usually compound potency against a given target. Only a subset of a series is displayed if analogues display variations at more than two substitution sites. Heat maps were also used to display mean potency changes resulting

from the exchange of a pair of substituents at a given site.⁵ Similarly to SAR maps, multiple views of the same series of analogues are required to display SAR information for more than one substitution site.

Another recently introduced data structure of graphical analogue analysis is the Combinatorial Analogue Graph⁶ (CAG) that systematically organizes substitution sites and their combinations in a tree-like structure and identifies SAR hotspots making large contributions to SAR discontinuity. Hence, CAGs view analogue series from a perspective different from that of R-group tables because they pinpoint substitution sites in the common core structure where R-groups make important contributions. However, CAGs do not provide an immediate access to functional groups at these positions.

Other than R-group tables, their extensions, and CAGs, there are currently no graphical SAR analysis methods for analogues available. In particular, SAR relationships between R-group combinations at different sites cannot be analyzed in a straightforward and consistent manner. Therefore, we have been interested in the development of a graphical data structure that goes beyond the capacities of previously published visualization methods by explicitly using R-group combinations and their (subset) relationships as an organizing principle.

Following our approach, R-group combinations are systematically extracted from a given analogue series, associated with potency information of all analogues containing a specific combination, and organized according to consistently numbered substitution sites. Importantly, subset relationships between R-group combinations emerge from this data structure such that

Received: October 10, 2011

Published: January 16, 2012

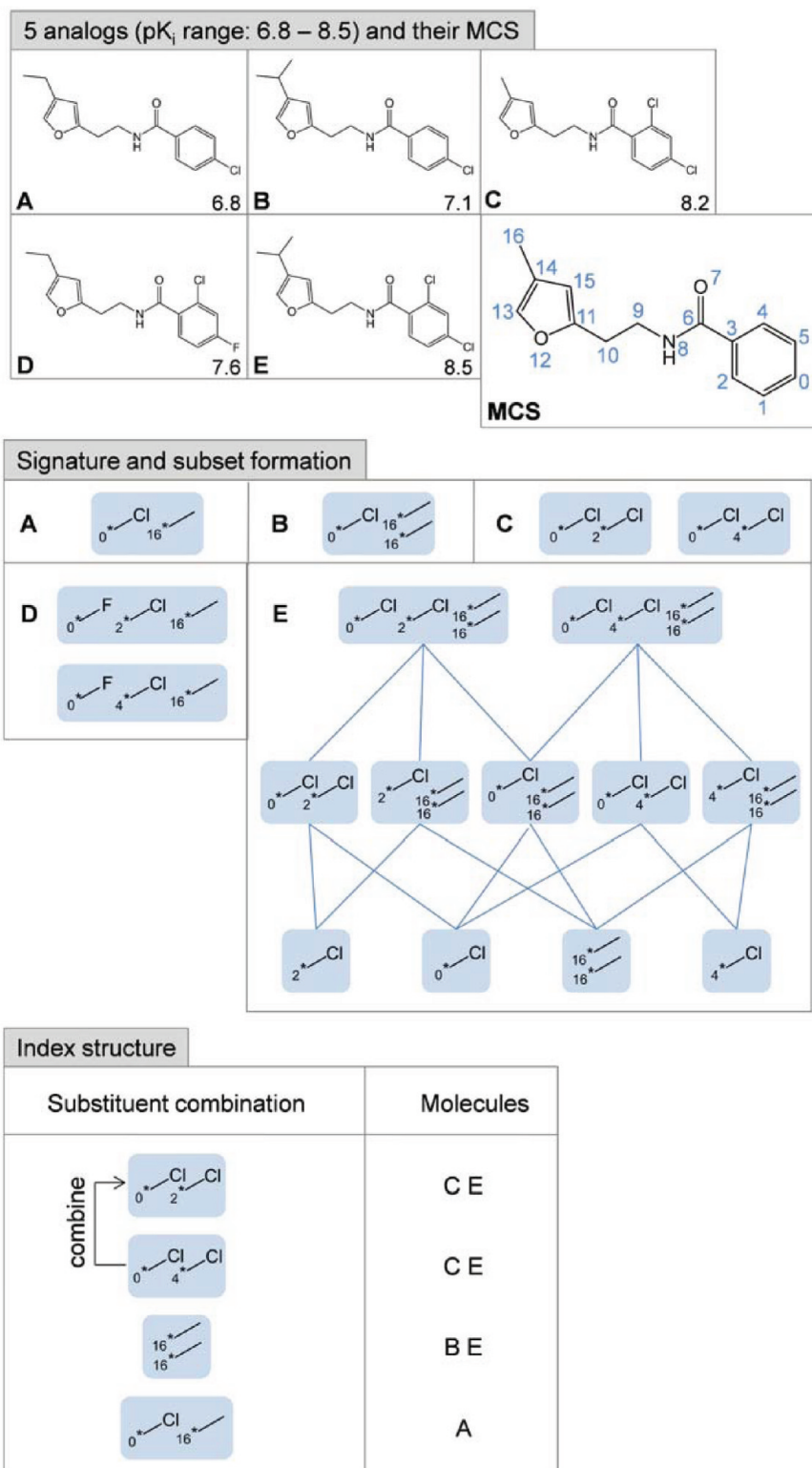


Figure 1. R-group combinations. A model analogue series of five compounds is shown (top) together with its maximum common substructure (MCS). Compounds A–E are annotated with their pK_i values. Signatures, i.e., sets of R-groups, are extracted from all compounds (middle). For molecule E, the generation of signature subsets is illustrated. R-group combinations are then added to an index structure and associated with the analogues in which they occur (bottom). For clarity, only a section of the complete index structure is shown. R-group combinations that are obtained by symmetry-related mappings (as explained in Methods) are combined into a single entry.

potency changes resulting from the removal (addition) of a substituent from (to) a given R-group combination can be

monitored. Our methodology is introduced herein, and exemplary applications are provided.

METHODS

R-Group Deconvolution and Signature Formation. Initially, the maximum common substructure (MCS) of compounds comprising an analogue series is determined, and all non-hydrogen atoms of the MCS are labeled with numeric identifiers, as illustrated in Figure 1 (top) for a model series of five analogues. The MCS is then used as the invariant molecular core structure for R-group deconvolution of all analogues. For this purpose, the MCS is mapped onto each analogue, and the numeric identifiers are transferred to matching atoms. Variable R-groups are identified and unambiguously assigned to corresponding substitution sites for all analogues by extracting groups that are not part of the alignment and marking them with the numeric identifier(s) of the matching atom(s) to which they are attached. The list of all identified R-groups is used as the *signature* of the molecule. If multiple mappings of the MCS onto a compound are possible because it contains symmetry elements around rotatable bonds, all mappings and resulting signatures are determined, as illustrated for molecules C, D, and E in Figure 1 (middle). In this example, the three molecules contain a chlorine atom at the ortho position of the phenyl ring that can be assigned to substitution sites R2 or R4.

In the next step, R-group combinations that are shared by multiple compounds are systematically detected by extracting signature subsets from all analogues. Hence, if an analogue contains R-groups at n substitution sites, all possible signature subsets with R-groups for $n - 1$ to 1 substitution sites are generated, as illustrated for molecule E in Figure 1 (middle). The original signature and all signature subsets are then added as separate keys to an index table and assigned to the source compound (Figure 1, bottom). Hence, all analogues belonging to a particular key share the R-group combination defined by the key.

If alternative mappings of the MCS onto given analogues of a series are possible, corresponding keys in the index might describe the same R-group pattern for an identical set of compounds with alternatively numbered substitution sites. These keys are identified and combined into a single entry (Figure 1, bottom). The R-group decomposition, signature (subset) formation, and index structure generation routines were implemented in Java using the OpenEye chemistry toolkit.⁷

Graph Design and Visualization. In order to capture subset relationships between keys in the index table (i.e., sets of R-groups at specific substitution sites), a *directed acyclic graph* is generated, as illustrated in Figure 2.

Graph Structure. Keys correspond to *nodes* in the graph. Each node is associated with the set of molecules that contain the specified R-group combination (and are thus linked to the same key in the index table). Nodes are connected via *directed edges* to all other nodes that are obtained by removing R-groups from exactly one substitution site of the original set. Thus, nodes connected by directed edges are involved in parent–child relationships, and all molecules that are associated with a parent node are also associated with a child node. However, a child node might contain additional analogues.

A child node associated with exactly the same set of molecules as its parent node is removed, which reduces the complexity of the graph by omitting redundant information. This is the case if a smaller R-group combination always occurs in the context of a larger one. If node removal eliminates the only existing pathway between a parent and a grandchild (i.e., another node connected to the child), an edge is inserted that directly connects the parent to its grandchild. The graph structure is iteratively updated after each node removal. The process ends when all redundancies are eliminated. The original unprocessed graph structure for the model analogue series in Figure 1 is shown in Figure 2a. Because all edges from the top to the bottom of the graph are directed (and follow the same direction), arrows are generally omitted for clarity. Nodes that convey redundant information and are removed from the graph during processing are shown in yellow. The processed graph is depicted in Figure 2b.

Node Types. In the processed graph, two types of nodes are distinguished: nodes that are associated with a single compound are drawn as *circles* while nodes that represent R-group combinations in multiple analogues are drawn as *squares*. The size of an analogue subset

assigned to a square-shaped node is indicated by its *frame thickness* that increases with the number of compounds.

The different node types are interpreted as follows: the R-group combination represented by a circle node corresponds to the signature (i.e., the complete list of R-groups) of the single molecule that is assigned to the node. Hence, the combination of the signature and MCS defines the molecular structure of the associated compound. However, the signature of a molecule is only associated with a circular node if the corresponding set of R-groups does not occur in any other analogue of the series. If the signature of a compound corresponds to a subset of R-groups in other analogues, these analogues are combined and represented by a square-shaped node. Following our terminology, this compound is then *masked* by the square-shaped node. In order to identify a masked compound in the graph, it is symbolized as a rectangle in the lower-right quadrant of the node (Figure 2b).

Compound Potency Information. All circle nodes are colored according to the potency of the corresponding compounds, and the square-shaped nodes are colored according to the mean potency of the associated analogues using a uniform continuous color gradient from green (lowest potency in the data set) to red (highest potency), as illustrated in Figure 2c. A rectangle symbolizing a masked compound is colored according to its potency (analogous to circle nodes). Square-shaped nodes are often not completely color-filled, for the following reason: if multiple compounds are associated with a node, the *area* of the node that is colored reflects the standard deviation of potency values. Thus, a node that is completely colored corresponds to a standard deviation of zero, i.e., all associated molecules have the same potency value. For standard deviations larger than zero and smaller than one, the color-filled area continually decreases to half of the original diameter and is then kept constant for standard deviations equal to or larger than one (Figure 2c). Hence, decreasing color-filled node areas indicate increasing compound potency variations.

As shown in Figure 2, nodes are arranged in layers that reflect decreasing numbers of substitution sites, i.e., parents are always positioned above their children. Furthermore, within the same layer, nodes representing R-groups at exactly the same substitution sites are grouped together and arranged in order of increasing potency from left to right.

Implementation. R-groups represented by nodes are stored as canonical SMILES⁸ strings. Nodes are associated with tooltips to display R-group structures and report the number of compounds assigned to a node, as well as their mean potency, and the potencies of any masked compounds. The graph layout can also be interactively edited. The graph design was implemented using the Java package JUNG.⁹ Because the graph structure emphasizes relationships between different R-group combinations, as discussed in detail in the following, it was termed *directed R-group combination* (DRC) graph (DRCG).

RESULTS AND DISCUSSION

The DRC graph structure is designed to extract SAR information from R-group patterns in analogue series. An important feature of the approach is that any series of analogues can be studied in context, regardless of the number of substitution sites that occur (or the number of compounds). Furthermore, the systematic and hierarchical organization of analogues on the basis of combinations of all R-groups that are available in a series and the analysis of relationships between different sets of R-groups also set this methodology apart from currently available approaches to study analogue series such as R-group tables and their extensions. In particular, the multiple R-group analysis scheme reveals (i) critical substitution sites, (ii) (un)favorable substituents, (iii) additive and nonadditive effects on compound potency as a consequence of multisite substitutions, (iv) optimization pathways gradually increasing compound potency, and (v) suggestions for analogue design. Thus, as shown in the following, the potential of the DRCG approach goes much beyond conventional analysis of analogue series.

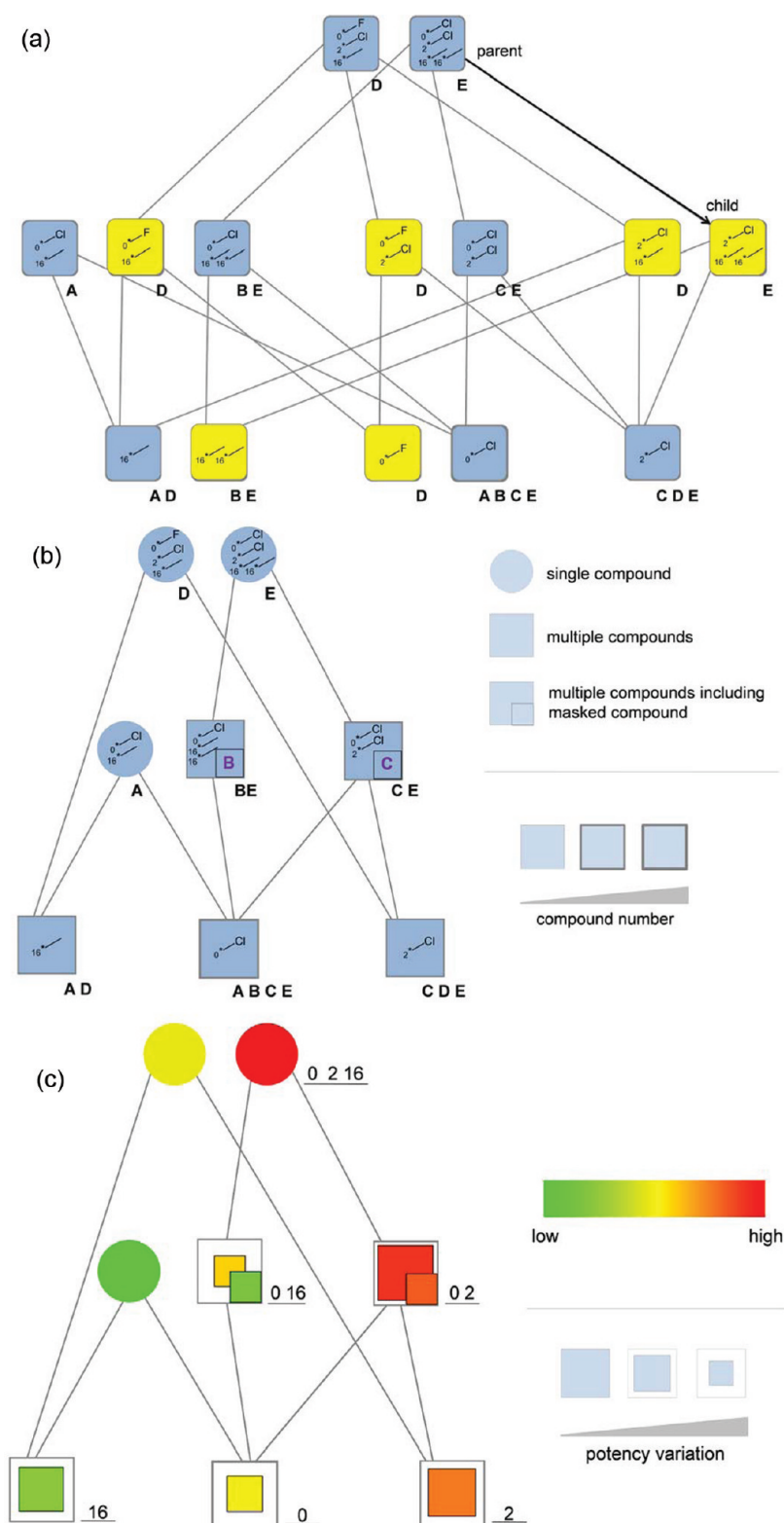


Figure 2. Graph structure. Schematic illustrations of the graph structure are presented that highlight different design elements. (a) An unprocessed graph is displayed that contains nodes for all substituent combinations found in the model data set shown in Figure 1. Nodes are associated with all analogues containing the given R-group combination. An exemplary parent–child relationship between two R-group sets and the corresponding directed edge are indicated on the right. Nodes that carry redundant information because they are associated with the same analogue subset as a parent node are highlighted in yellow. (b) The processed graph is shown after (i) removal of redundant nodes, (ii) introduction of different node types, and (iii) scaling of node frames according to compound numbers. For two nodes, the masked compounds B and C are labeled in purple. (c) The graph is displayed with (i) a color code accounting for (mean) compound potencies and (ii) scaling of color-filled node areas according to potency variations. For clarity, node labels and compound information are not shown. Instead, groups of nodes are labeled with the corresponding substitution site combinations.

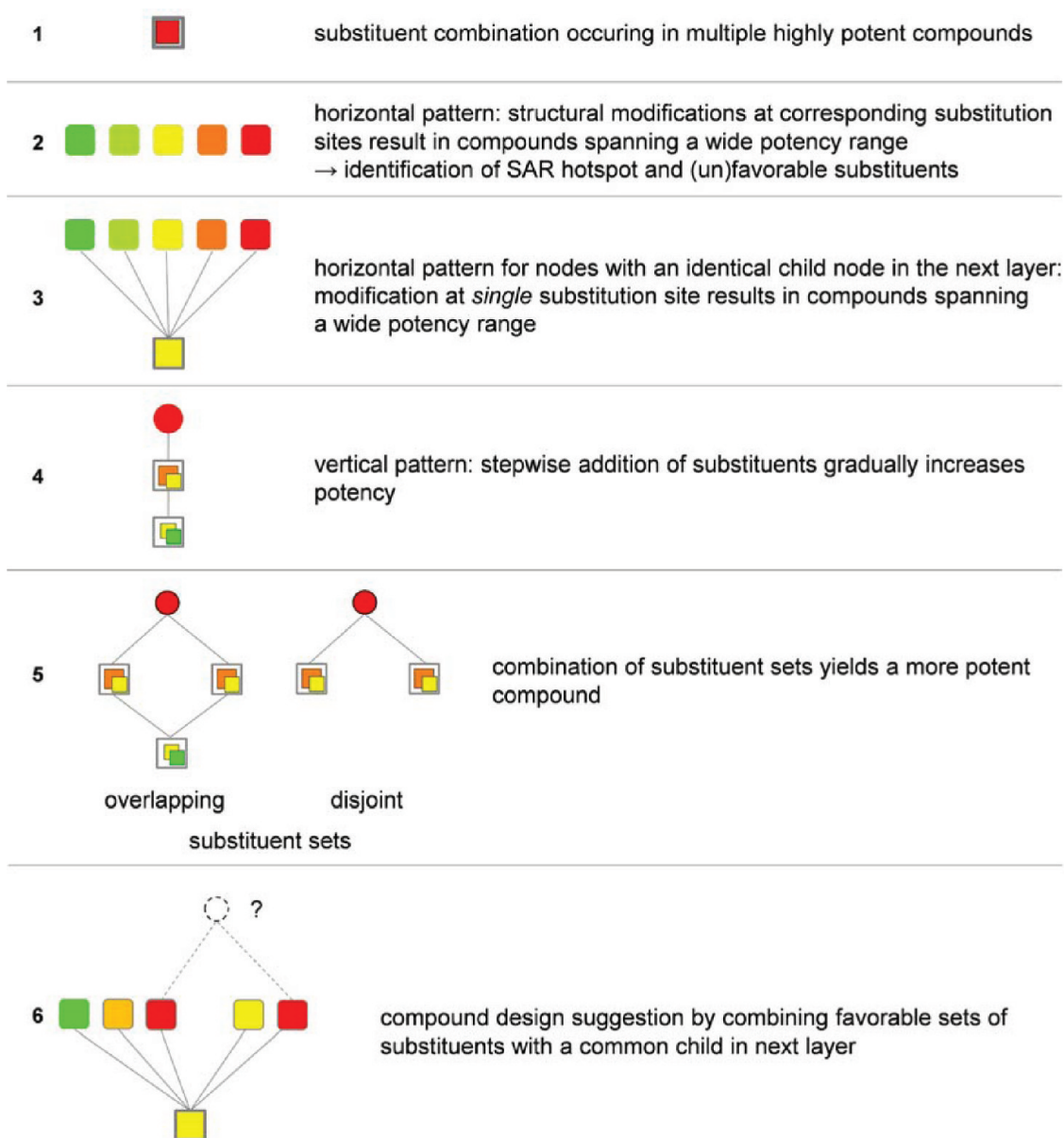


Figure 3. SAR patterns. Node patterns that represent characteristic features of the graph representation and capture SAR information in a defined manner are shown and explained.

SAR Patterns. Following our approach, it is of central importance that interpretable SAR information is readily extracted from a graph structure. The DRCG structure contains several well-defined subgraphs that reveal immediately interpretable SAR information. These graph components are termed *SAR patterns* and schematically depicted in Figure 3. These patterns are rationalized as follows:

SAR pattern 1: This is the formally simplest pattern. R-group combinations that exclusively occur in highly potent compounds are identified by square-shaped nodes filled with red color that, ideally, have a thick frame indicating that the R-group combination has been explored in many different compounds that are consistently highly potent.

SAR pattern 2: Critical substitution sites or combinations of sites where structural modifications lead to

large differences in potency occur as *horizontal* node patterns in the graph. In this case, differently colored nodes are grouped together within the same node layer, hence representing different combinations of R-groups at the same substitution sites spanning a wide potency range. It follows that this pattern also provides an immediate access to favorable and unfavorable R-group combinations.

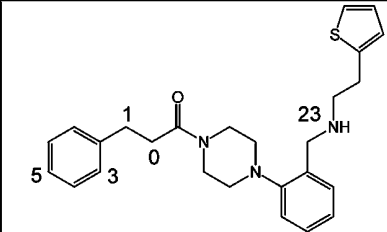
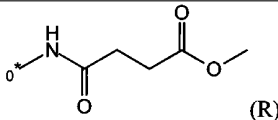


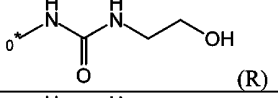


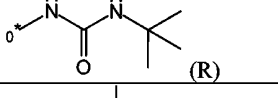


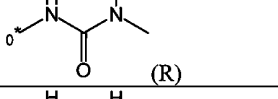


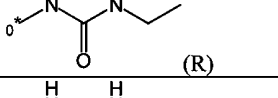


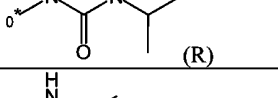


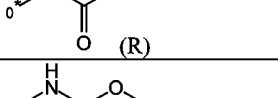


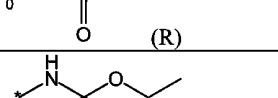


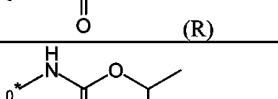


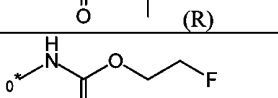


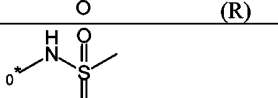


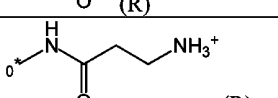


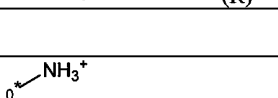






SAR pattern 3: If the nodes forming the horizontal pattern 2 are all connected to the same child node in the subsequent layer, the structural modifications responsible for large potency variations can be traced back to a single substitution site.

SAR pattern 4: A gradual increase in potency resulting from a stepwise addition of R-groups to a starting compound is detected as *vertical* pattern. Following the path from a node in inverse edge direction

Table 1. R-Group Table Representation of an Analogue Series^a

R0	R1	R3	R5	R23	pK _i
					6.0
					6.2
					5.2
					7.2
					7.4
					6.5
					7.7
					7.9
					7.4
					7.2
					7.5
					6.4
					7.0
					7.7
					7.6
					8.7
					8.7

Table 1. Continued

					
R0	R1	R3	R5	R23	pK _i
 (R)					8.6
 (R)					8.5
 (R)					8.6
 (R)					8.7
 (R)					8.8
 (R)					8.9
 (R)					8.6
 (R)					8.3
 (R)					8.3
 (R)					8.3
 (R)					8.5
 (R)					7.9
 (R)					6.7
					5.3
					6.0

^aFor a series of 32 antagonists of the melanocortin receptor 4, the common core structure and substitution sites are provided in a conventional R-group table format. For all individual analogues, R-groups and potency values are reported. For a subset of analogues, the stereocenter at substitution site R0 is in the R-configuration as indicated in the table (R).

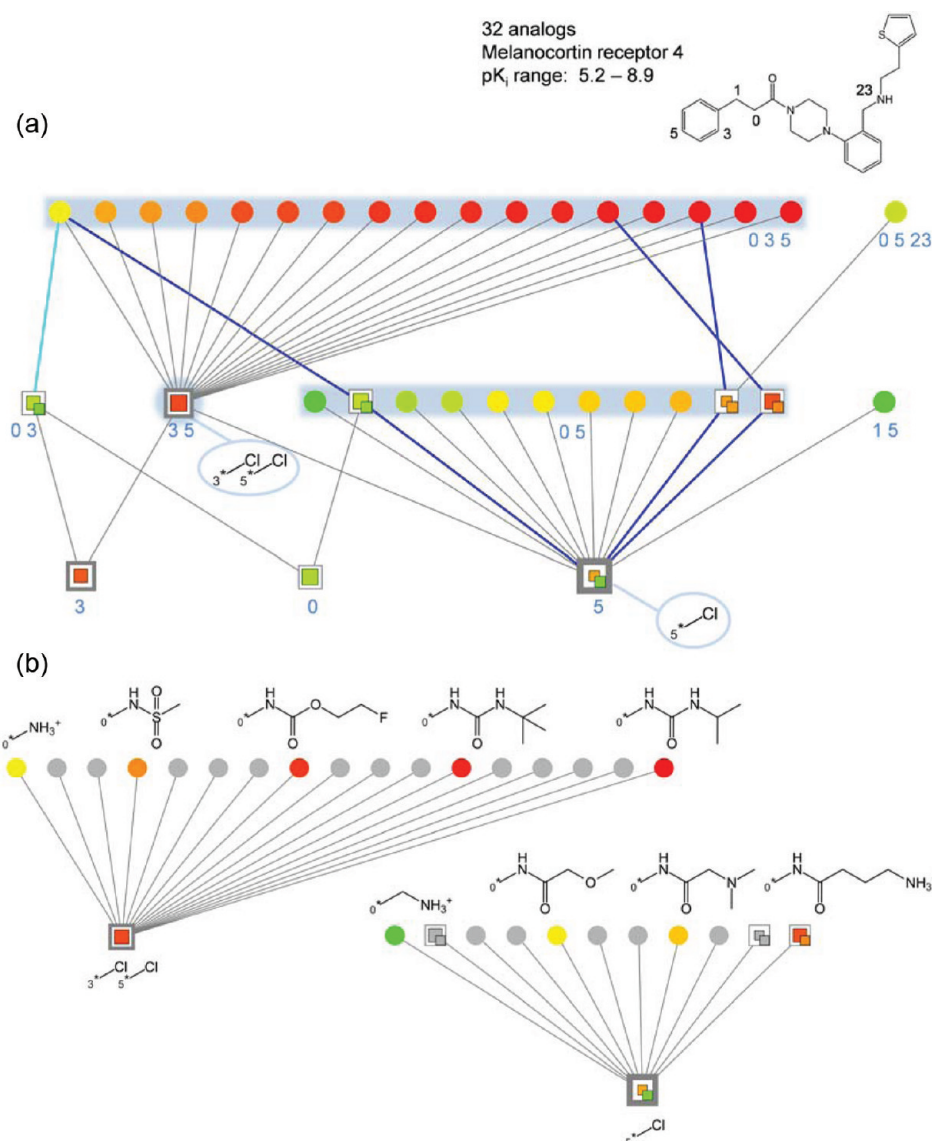


Figure 4. Melanocortin receptor 4 antagonists: series 1. (a) The MCS of a series of 32 analogues is shown at the top, and substitution sites are labeled with numeric atom identifiers. For two nodes, substituent combinations are provided. Characteristic SAR patterns are numbered according to Figure 3 and highlighted as follows: SAR patterns 1 and 3, blue node background; pattern 4, blue edges (on the right); pattern 5, combination of blue and turquoise edges (left). (b) Exemplary analogues from horizontal patterns are shown. The remaining nodes are colored gray.

(i.e., bottom-up toward its ancestors) leads to increasingly potent analogues.

SAR pattern 5: A parent node that is connected to multiple less potent child nodes indicates that its potency results from the interplay of the different R-group sets associated with the child nodes. The substituent sets of the child nodes are disjoint unless they share a common ancestor. The interplay between different substitution sites and R-groups might result in additive or nonadditive effects on compound potency.

SAR pattern 6: Under the likely assumption that favorable R-group effects on compound potency are not compensatory (i.e., that positive effects at two or more sites do not combine in a negative way), compound design suggestions can be easily made on the basis of the DRCG structure. Attractive analogues with predicted high potency can be

derived from nodes that represent favorable R-group combinations within the same layer and are connected to a shared, less potent child node in the next layer. Thus, starting from the same R-group combination, the introduction of additional R-groups at different substitution sites leads to analogues with increased potency. It follows that new analogues can be immediately suggested that combine the original R-group set with all potency-increasing R-groups introduced at distinct sites.

If SAR information is contained in a series of analogues, it will consistently emerge in the form of the intuitive SAR patterns described above. Therefore, searching a DRCG of any analogue series for these characteristic SAR patterns enables the extraction of SAR information, if available in a data set.

Exemplary Applications. We applied the DRCG method to four analogue series of different composition containing

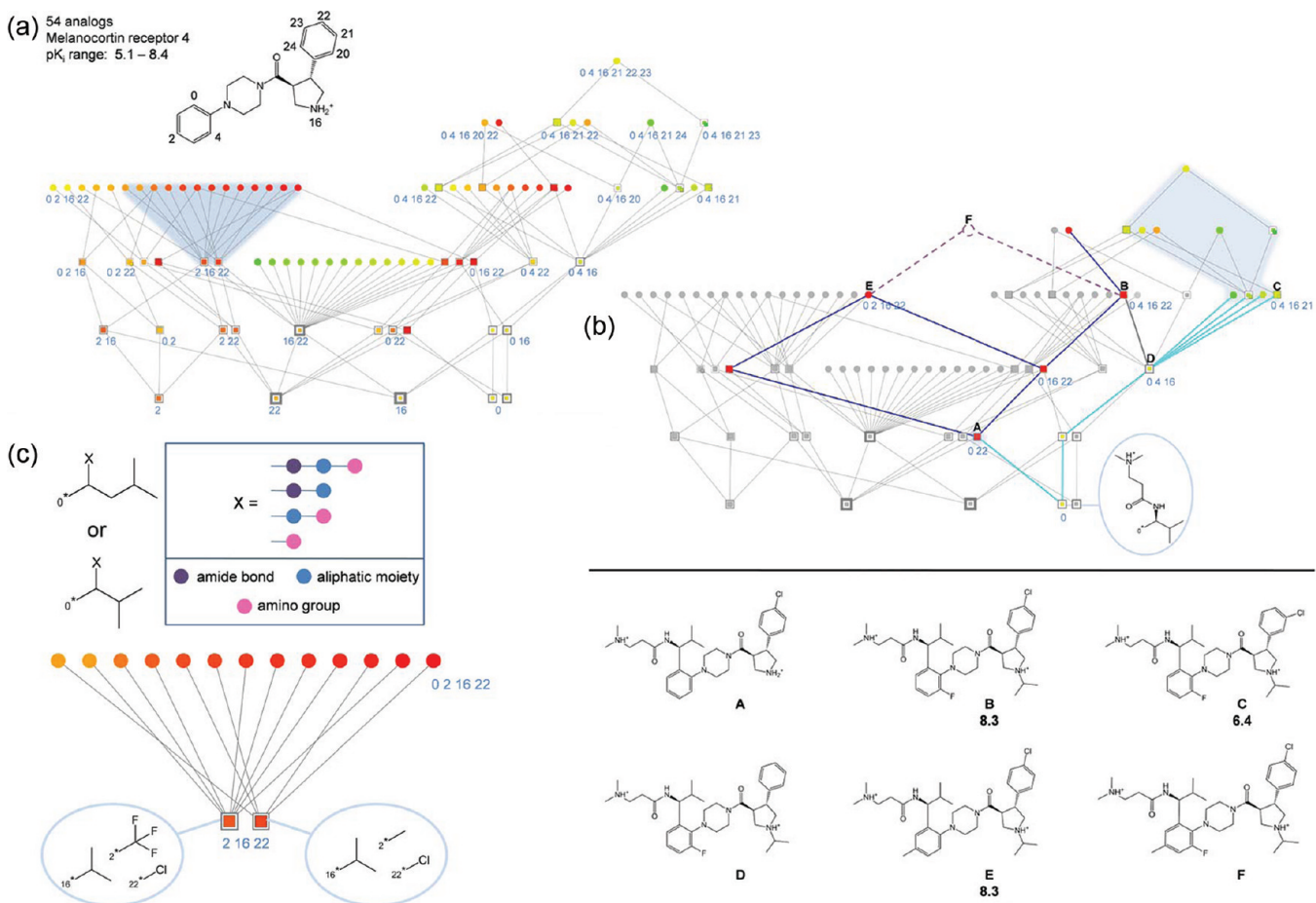


Figure 5. Melanocortin receptor 4 antagonists: series 2. For a series of 54 analogues, the complete graph and SAR information-rich subgraphs are shown. (a) The MCS of all analogues is shown with relevant numeric atom identifiers. A subgraph associated with highly potent compounds and two substituent combinations frequently found in these analogues are highlighted (SAR pattern 1, blue background). (b) Starting from a favorable combination of two R-groups at substitution sites R0 and R22 (SAR pattern 1), edges leading to highly potent compounds containing this combination are shown in blue. A cluster of weakly potent compounds is highlighted (blue background) that contains only one of these two R-groups (at R0) and additional R-groups at other sites. The path to this cluster is indicated using turquoise edges. A compound design suggestion is indicated by dashed magenta edges. (c) The subgraph corresponding to the highlighted SAR pattern in (a) is shown together with R-group information for nodes.

between 31 and 54 compounds. Initially, compounds active against the human melanocortin receptor 4 (MC4R), norepinephrine transporter (NET), and dopamine D1 receptor (DRD1) were extracted from the ChEMBL¹⁰ database. From compounds forming each activity class, Bemis and Murcko scaffolds¹¹ were extracted and molecules sharing the same scaffold (and activity) were combined into an analogue set. Series comprising 30 or more analogues were subjected to DRCG analysis. In the following, representative examples of series are discussed that contain interpretable SAR information.

Melanocortin Receptor 4 Antagonists. Analogue Series 1. The first series of MC4R antagonists consists of 32 analogues with potencies ranging from pK_i 5.2 to 8.9. For this series, a conventional R-group table is provided in Table 1, and its DRCG representation is shown in Figure 4a. Because analogues sharing the same substitution sites are grouped together, the graph reveals that most analogues in this series are characterized by two different substitution site combinations, i.e., R0/R5 (lower horizontal pattern in Figure 4a) and R0/R3/R5 (upper horizontal pattern). In both cases, R-groups at site R0 vary whereas groups at site R5 and sites R3/R5 are invariant. This is captured by the graph structure because

removal of the substituents at R0 yields the same child for all nodes of a group, i.e., child nodes annotated with “3 5” and “5”, respectively. The labels of these nodes reveal that the invariant R-groups at R3 and R5 are chlorine atoms. Analogues forming both horizontal patterns are arranged in order of increasing potency. In both instances, traversing nodes and associated R-groups from the left to the right reveals that aliphatic amine moieties attached to R0 via an amide bond or carbamide derivatives are preferred substituents. Exemplary analogues are depicted in Figure 4b.

Because R-groups at R0 are highly variable, only three of the analogues in the upper horizontal pattern are derived from others by addition of a chlorine atom to R3. However, from the vertical pathways involving these analogue pairs (highlighted in blue in Figure 4) it can be inferred that a chlorine atom at R3 increases compound potency. The leftmost compound in the upper horizontal pattern is accessible via two edges because an analogue that differs from this compound by the absence of the chlorine substituent at R5 is also available in the data set. The analogue without the chlorine at R5 is also less potent: we can thus conclude that chlorine atoms at R3 and also R5 make positive contributions to compound potency, which is also reflected by the framed red-filled “3 5” node representing this R-group pair.

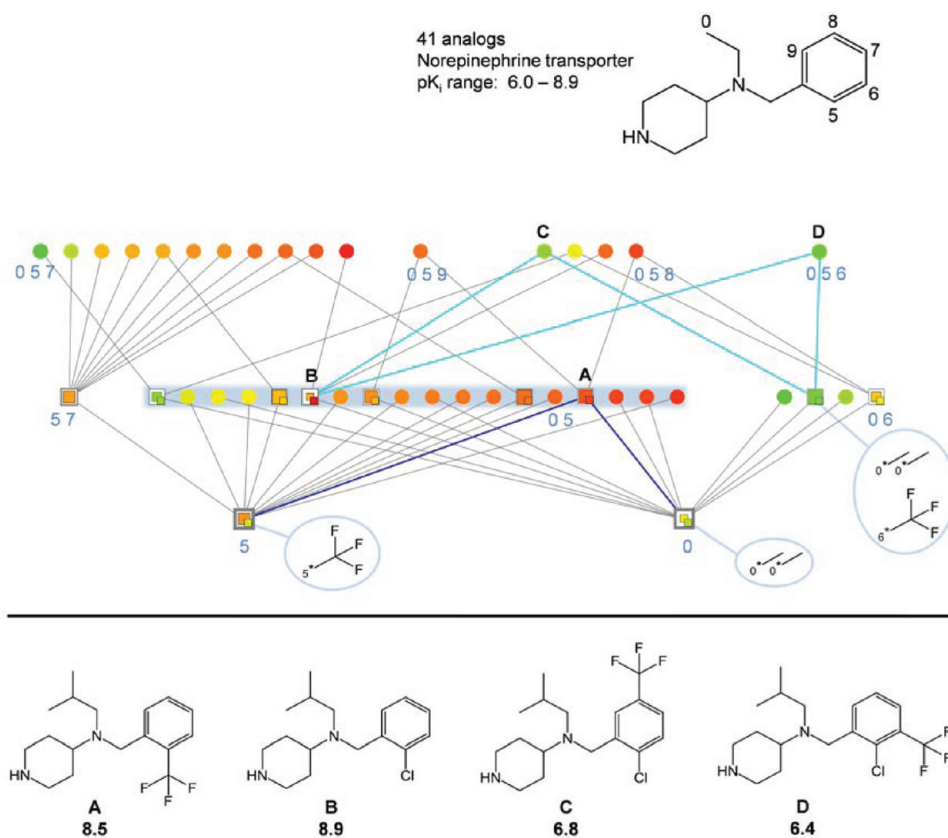


Figure 6. Norepinephrine transporter inhibitors. The MCS extracted from a series of 42 analogues is shown with numeric atom identifiers. For three nodes, substituent combinations are shown. In addition, four nodes are labeled with the identifiers of compounds (A–D) defined by the corresponding substituent combinations. Structures and potency information for these compounds are provided at the bottom. SAR patterns are highlighted: pattern 2, blue node background; SAR pattern 5, blue edges. In addition, detrimental effects of R-group combinations on compound potency are indicated using turquoise edges.

By comparison with Table 1 it is evident that the DRCG structure provides easy access to SAR information contained in this series that would be much harder to extract from an R-group table. In particular, the analysis of multisite R-group effects would require a comparison of an analogue with all others in the R-group table. Furthermore, in the graph structure, a set of R-groups is associated not only with the potency of the analogue it defines but also with the mean potency of all compounds in which this R-group combination occurs. For example, the node that represents chlorine at position R5 in Figure 4 (bottom right) provides the information that this substitution alone (masked in this node) yields only a weakly potent compound, whereas its combination with R-groups at other sites generally produces compounds with increased potency. This type of information conveyed by the DRCG representation helps to identify R-groups that are favorably in combination with others.

Analogue Series 2. The DRCG of another structurally distinct series of 54 MC4R antagonists covering a pK_i range from 5.1 to 8.4 is shown in Figure 5a. The complete graph reveals that analogues in this series are substituted at three or more different sites. A region formed by increasingly potent analogues is highlighted. In these analogues, two different combinations of R-groups at substitution sites 2, 16, and 22 are frequently found (corresponding to the two nodes with the “2 16 22” label at the bottom of the highlighted pattern). Figure 5b focuses on another combination that consistently produces highly potent analogues. This combination is formed by a 2-methylpropyl-3-(dimethylamino)propanamide R-group at

position R0 and a chlorine atom at R22. Analogue A in Figure 5b that contains only these two substituents is currently untested (as stated above, all known analogues comprising this series carry R-groups at three or more sites). The hypothetical analogue A would also be associated with the “0 22” node. Edges forming pathways to all compounds that contain this site combination are highlighted in blue in Figure 5b. Furthermore, all analogues that contain the same R-group at R0 but lack the chlorine substituent at R22 (corresponding to the “0” node) are reached following the turquoise edges, leading to a cluster of weakly potent compounds highlighted on the right of the graph. Two analogues B and C are also marked in the graph that contain the 2-methylpropyl-3-(dimethylamino)propanamide group at R0 and identical R-groups at R16 and R4, as indicated by an additional chlorine atom at R22 and is highly potent whereas the other contains a chlorine atom at R21 and is only weakly potent. Therefore, it would be suggested to test another hypothetical compound D (also shown in Figure 5b) that does not contain any of the chlorine substituents, which might confirm the potency-decreasing effect of a chlorine atom at R21 and/or the potency-increasing effect of a chlorine atom at R22.

Overall, most potent compounds are obtained when substitution sites R0, R2, R16, and R22 are simultaneously occupied, as highlighted in Figure 5a. Many analogues with this site combination contain an isopropyl group at R16, a chlorine atom at R22, and a methyl or trifluoromethyl group at R2, as shown in Figure 5c. The R-group at R0 is generally large and

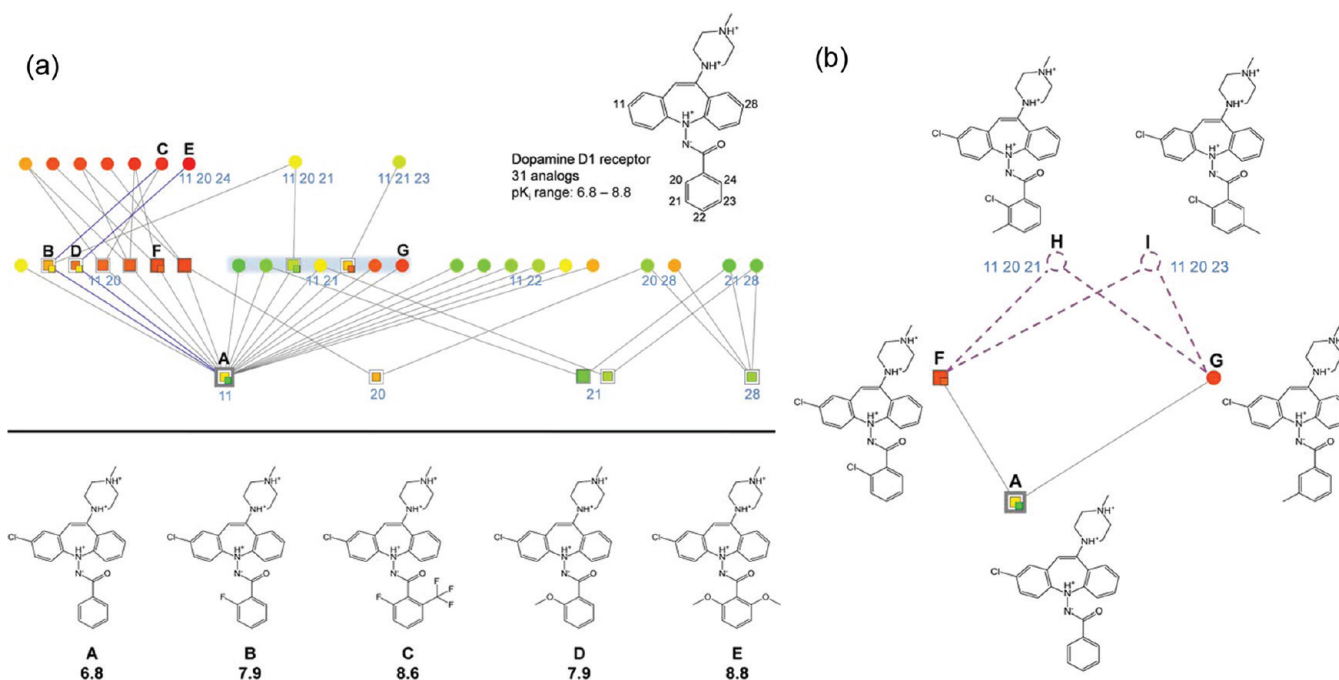


Figure 7. Dopamine D1 receptor antagonists. For a series of 31 analogues, the complete graph and a subgraph illustrating compound design suggestions are shown. (a) The MCS with relevant numeric atom identifiers is shown. Characteristic SAR patterns are highlighted: pattern 3, blue node background; pattern 4, blue edges. (b) Compound design suggestions (H, I) based on SAR pattern 6.

has limited structural variability, as illustrated at the top of Figure 5c, which displays the subgraph associated with these highly potent analogues. Compounds associated with nodes B and E in Figure 5b share the same R-groups at overlapping substitution sites (R0/R16/R22; shared child node) and have consistently high potency. Hence, it would be interesting to combine these favorable R-group sets by generating a new analogue F that is shown in Figure 5b.

Norepinephrine Transporter Inhibitors. The DRCG of a series of 41 NET inhibitors spanning a potency range of approximately three orders of magnitude is shown in Figure 6. The graph representation reveals that R5 and R0 have predominantly been explored in this series. Recurrent R-groups among analogues include a trifluoromethyl group at R5 and a dimethyl rest at R0. Interestingly, the introduction of one of these R-groups in isolation only generates a weakly potent compound, but their simultaneous introduction leads to a more than additive increase in potency, yielding one of the most potent analogues in this series (masked in node A; labeled A in Figure 6). The highlighted horizontal pattern for the combination of substitution sites R0 and R5 shows that the introduction of different R-groups at both sites has large effects on compound potency. Moreover, as revealed by the weakly potent compounds on the right in Figure 6, the introduction of additional substituents at the meta positions of the phenyl ring displays a strong tendency to decrease potency. For example, when adding a trifluoromethyl group to the most potent analogue of this series (B in Figure 6) at R8 (yielding analogue C) or R6 (D), potency is reduced by more than 2 orders of magnitude. In general, the relationships between di- and trisubstituted analogues in the DRCG of this series indicate that the addition of R-groups at sites other than R0 and R5 does not lead to notable increases in potency.

Dopamine D1 Receptor Antagonists. The DRCG of a series of 31 DRD1 antagonists that span a comparably narrow pK_i range from 6.8 to 8.8 is shown in Figure 7a. As revealed by

the highlighted horizontal pattern, the introduction of different chemical groups at the meta position of the terminal phenyl ring (designated R21 in the graph) leads to largest potency fluctuations within this series. At the left and right of this pattern, the trifluoromethyl and methyl group are identified as least and most favorable R-group at this site, respectively. In addition, considerable potency increases are observed for all analogues having an R-group at the ortho position of the phenyl ring (designated R20). Two vertical patterns are highlighted in Figure 7a where the subsequent addition of R-groups leads to stepwise increases in potency. Both pathways begin at the same analogue that carries a chlorine substituent at R11 and is only moderately potent (analogue A). The addition of a fluorine atom at R20 then leads to a potency increase of approximately 1 order of magnitude (analogue B). Another order of magnitude is gained by adding a trifluoromethyl group at the other ortho position in the ring (R24, analogue C). Similar potency changes are detected for the stepwise addition of two methoxy groups at the corresponding positions (analogues D and E). However, as depicted in more detail in Figure 7b, potency changes of larger magnitude are observed for compounds F and G that are also derived from analogue A. In these cases, both the introduction of a chlorine atom at the ortho position or a methyl group at the meta position of the terminal phenyl ring increase compound potency by 2 orders of magnitude. Hence, it would be attractive suggesting two additional analogues that combine these favorable substitutions (i.e., hypothetical molecules H and I in Figure 7b) in order to further increase compound potency within this series.

CONCLUSIONS

Herein we have introduced a new graphical SAR analysis concept specifically developed for the study of analogue series and for compound design. Instead of individual compounds, systematically derived R-group combinations provide the basis for the construction

of the DRCG structure, which is a central feature of the approach. The graphical representation contains a number of design elements that emphasize available SAR information. From the hierarchical organization of R-group combinations and corresponding analogue sets, characteristic subgraphs emerge that represent well-defined SAR patterns. If analogue series are characterized by the presence of multiple substitution sites and R-group combinations, it is usually difficult to rationalize SARs by comparing individual analogues and their potency values in R-group tables or subset of analogues with pairs of substitution sites. By contrast, in DRCGs, entire analogue series are consistently represented, regardless of the numbers of analogues and substitution sites, and emerging SAR patterns reveal interpretable SAR information. In this study, we have presented the design of the DRCG structure and discussed characteristic features of the graph structure. In addition, exemplary analyses of different analogue series have been carried out to illustrate how SAR determinants are identified on the basis of interactive graphical analysis and how such insights can be utilized to design new analogues.

Furthermore, the data structure can readily be extended to the analysis of compound properties other than potency. The hierarchical organization of substituent combinations and the display of subset relationships are central to the approach. However, properties encoded by node color and color-filled area are variables, and therefore the DRCG approach is easily adjustable to other applications. For example, node color could be used to encode a predominant mechanism of action (e.g., agonism, antagonism, partial agonism, or inverse agonism) of receptor ligands containing a given substituent combination. Moreover, compound potency might be replaced by selectivity in the DRCG calculation if SARs for two targets are analyzed in parallel. Hence, we anticipate that the DRCG approach introduced herein might become a valuable asset for the comprehensive analysis of multiple properties associated with analogue series.

We are currently in the process of further extending the DRCG approach to multiproperty analysis of analogue series, as outlined above. Upon completion of this extension, the method will be made available. Interested readers are referred to future updates on our homepage (see <http://www.lifescienceinformatics.uni-bonn.de>).

AUTHOR INFORMATION

Corresponding Author

*Tel: +49-228-2699-306. Fax: +49-228-2699-341. E-mail: bajorath@bit.uni-bonn.de.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank Mathias Wawer for helpful discussions. A.M.W. is supported by Boehringer Ingelheim.

ABBREVIATIONS:

SAR, structure–activity relationship; CAG, combinatorial analogue graph; DRC, directed R-group combination; DRCG, directed R-group combination graph; MCS, maximum common substructure; MC4R, melanocortin receptor 4; NET, norepinephrine transporter; DRD1, dopamine D1 receptor

REFERENCES

- (1) Wassermann, A. M.; Wawer, M.; Bajorath, J. Activity Landscape Representations for Structure–Activity Relationship Analysis. *J. Med. Chem.* **2010**, *53*, 8209–8223.
- (2) Ahlberg, C. Visual Exploration of HTS Databases: Bridging the Gap between Chemistry and Biology. *Drug Discovery Today* **1999**, *4*, 270–485.
- (3) Wawer, M.; Lounkine, E.; Wassermann, A. M.; Bajorath, J. Data Structures and Computational Tools for the Extraction of SAR Information From Large Compound Sets. *Drug Discovery Today* **2010**, *15*, 630–639.
- (4) Agrafiotis, D. K.; Shemanarev, M.; Connolly, P. J.; Farnum, M.; Lobanov, V. S. SAR Maps: A New SAR Visualization Technique for Medicinal Chemists. *J. Med. Chem.* **2007**, *50*, 5926–5937.
- (5) Agrafiotis, D. K.; Wiener, J. J. M.; Skalkin, A.; Kolpak, J. Single R-Group Polymorphisms (SRPs) and R-Cliffs: An Intuitive Framework for Analyzing and Visualizing Activity Cliffs in a Single Analog Series. *J. Chem. Inf. Model.* **2011**, *51*, 1122–1131.
- (6) Peltason, L.; Weskamp, N.; Teckentrup, A.; Bajorath, J. Exploration of Structure–Activity Relationship Determinants in Analogue Series. *J. Med. Chem.* **2009**, *52*, 3212–3224.
- (7) OEChem TK version 1.7.4.3; OpenEye Scientific Software Inc., Santa Fe, NM, 2010.
- (8) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.
- (9) Java Universal Network/Graph Framework version 2.0.1; <http://jung.sourceforge.net/> (accessed July 4, 2011)
- (10) ChEMBLDB. <http://www.ebi.ac.uk/chembl/> (accessed July 28, 2011)
- (11) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.