Exploration of Structure−Activity Relationship Determinants in Analogue Series

Lisa Peltason,[†] Nils Weskamp,[‡] Andreas Teckentrup,[‡] 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, Boehringer Ingelheim Pharma GmbH & Co. KG, Department of Lead Discovery, D-88397 Biberach/Riss, Germany

Received January 27, 2009

A computational methodology is introduced to systematically organize compound analogue series according to substitution sites and identify combinations of sites that determine structure−activity relationships (SARs) and make large contributions to SAR discontinuity. These sites are prime targets for further chemical modification. The approach involves the analysis of substitution patterns in “combinatorial analogue graphs” (CAG) and the application of an SAR analysis function to evaluate contributions of variable R-groups. It is applicable to analogue series spanning different potency ranges, for example, analogues taken from lead optimization programs or screening data sets (where potency differences might be subtle). In addition to determining key substitution patterns that cause significant SAR discontinuity, CAG analysis also identifies “SAR holes”, i.e., nonexplored combinations of substitution sites, and SAR regions that are under-sampled in analogue series.

Introduction

To explore SAR[a] information in hit-to-lead or lead optimization projects or extract SAR information from biological screening data, one typically studies series of active analogues. In this context, computational tools are often utilized to aid in the analysis of SAR features.[1] However, the applicability of computational models is often limited because SARs and the underlying activity landscapes can be highly variable for different compound classes.[2] An activity landscape is best rationalized as the biological response surface to changes in chemical structure. Thus, if one envisions a two-dimensional projection of chemical space with compound potency added as a third dimension, maps of varying topology are obtained that are reminiscent of geographical maps and reflect different SAR characteristics. For example, if progressive changes in compound structure only have small to moderate effects on biological activity, smooth activity landscapes are observed. By contrast, if changes in molecular structure result in substantial changes in potency, the resulting landscapes are rugged and include activity cliffs. In activity cliff regions, small changes in structure lead to dramatic (positive or negative) biological effects.[2] Hence, what might often appear to be outliers in compound series taken from screening data or errors in experimental measurements, both of which negatively affect computational SAR analysis,[1] might potentially represent some of the most interesting compounds for hit-to-lead projects because they indicate the presence of steep activity cliffs.[2] Analyzing SAR information on the basis of screening data is often complicated by the presence of many weakly active compounds and relatively narrow potency distributions of active analogues, if they exist. Accordingly, it is often difficult to establish SARs on the basis of these data and methods to aid in this process are yet to be developed.

However, regardless of compound data sources, variable relationships between molecular structure and biological activity substantially complicate our understanding of molecular similarity[3−5] and the ability to analyze and predict SARs. Given the often highly variable nature of activity landscapes and corresponding SARs and, in addition, the current inability to predict the strong compound class-dependence of SAR features,[5] it is not surprising that there continues to be significant interest in the development of experimental or computational approaches that aid in SAR analysis.

From a computational perspective, progress has recently been made through the introduction of SAR analysis functions that systematically relate compound similarity and potency to each other and quantify SAR features. These methods make it possible to study SARs on a large scale,[6,7] determine global SAR features,[6] identify activity cliffs,[6,7] or describe compound subsets that are related by different local SARs.[8] Combined with molecular network representations, SAR analysis functions have revealed local SAR features within compound data sets[7,8] that would be difficult to describe by other means.

We have been interested in systematically exploring SARs of analogue series taken from lead optimization or screening data in order to identify key substitution patterns that determine their SAR characteristics. This is a challenging task because methods are required that analyze SARs at the level of individual substitution sites and that must often be sensitive to relatively minor differences in potency. Therefore, we have organized compound series in what we call “combinatorial analogue graphs” (CAG) that present an easily understandable hierarchy of substitution patterns. These graph representations are annotated with local SAR Index (SARI)[8] scores to account for SAR discontinuity at the level of functional groups. Thus, CAG-SARI analysis makes it possible to systematically quantify SAR contributions of substitution sites and site combinations, graphically organize this information, and identify SAR hotspots and undersampled regions. Key substitution patterns can be identified

[*] To whom correspondence should be addressed: Phone: +49-228-2699-306, Fax: +49-228-2699-341, E-mail: bajorath@bit.uni-bonn.de.

[†] Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität.

[‡] Boehringer Ingelheim Pharma GmbH & Co. KG, Department of Lead Discovery.

[a] Abbreviations: CAG, combinatorial analogue graph; SAR, structure−activity relationship; SARI, structure−activity relationship index; QSAR, quantitative structure−activity relationship; Tc, Tanimoto coefficient.

10.1021/jm900107b CCC: $40.75
Published on Web 04/27/2009

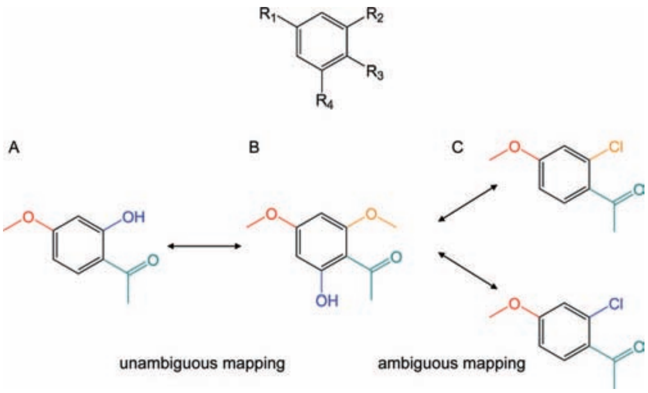


Figure 1. Substitution site assignment. Three analogues sharing a benzene framework are shown. The framework is substituted at up to four different substitution sites and can be mapped to each of the compounds in two different ways. For the symmetrically substituted compound B, mapping is unambiguous through the assignment of identical substituents in molecules A and B to corresponding sites. For compound C, the framework cannot be mapped unambiguously due to the presence of the chlorine substituent. This group is then arbitrarily mapped to one of the two possible substitution sites in molecule B.

and subsets of analogues that are most relevant for optimization efforts. Our methodology and practical applications are reported herein.

Materials and Methods

SARI Discontinuity Score. The discontinuity score of the SARI function[6] is used to quantify SAR features within given compound series. SAR discontinuity is characterized by large changes in potency as a consequence of small chemical modifications. Hence, the discontinuity score measures high potency differences among similar compounds and thus accounts for the presence of activity cliffs within a set of active molecules. For each pair of molecules that exceed a predefined similarity threshold, the product of their potency difference and their pairwise similarity is calculated. The discontinuity score for a set of active molecules is then calculated as the average of the product of pairwise potency difference and similarity for all similar compound pairs. Thus, the discontinuity score is defined as the average potency difference among similar compound pairs, scaled by pairwise similarity in order to emphasize potency differences between highly similar compound pairs:

$$\text{disc} = \underset{\{(i,j)|\text{sim}(i,j)>0.65,i\neq j\}}{\text{mean}} (|P_i - P_j| \times \text{sim}(i, j))$$

Here P_i and P_j denote the potency values of compounds i and j, and sim(i,j) denotes their similarity, calculated simply as the MACCS[9] Tanimoto coefficient (Tc).[10] We calculate discontinuity scores on the basis of compound subsets from analogue series that differ only at specific sites, as described below, i.e., high scores indicate subsets of compounds that include significant activity cliffs. The “raw” (i.e., non-normalized) discontinuity score is normalized with respect to the raw scores of all compound subsets from all analogue series in a given data set. As described previously,[8] all score values within a data set are used as reference to calculate z-scores for the discontinuity scores. The z-scores are then mapped to the value range [0,1] by calculating the cumulative probability for each score under the assumption of a normal distribution. Thus, score distributions are characteristic of a given data set and hence also make it possible to differentiate relatively narrow potency distributions.

Similarity Assessment. As stated above, the pairwise similarity between two molecules is assessed as the Tc value calculated for MACCS structural keys that have originally been developed for substructure mapping. While a wealth of other similarity metrics and more complex descriptors of chemical structure and properties exist, MACCS keys are found to produce chemically meaningful and easily interpretable results. However, in general, the methodology presented herein can be applied using any chemical similarity measure.

Analogue Series Identification. We automatically extract series of analogue structures from source data sets (Table 1) through analysis of molecular frameworks following Bemis and Murcko.[11] Accordingly, core structures are derived by deleting all R-groups from a molecule and rings and linkers are retained together with atom element, hybridization, and bond order information. Molecules with identical frameworks are then grouped into analogue series. For this study, large analogue series of up to approximately 100 compounds were selected in order to provide a meaningful basis for score calculations.

R-Group Decomposition. Compounds in analogue series are divided into constant and variable regions through R-group decomposition. Typically, invariant regions include the molecular framework and possibly R-groups that are conserved in all compounds of a series. Initially, invariant molecular regions are determined by calculating the maximum common substructure (MCS) of all analogues in a series. The MCS is then used as core structure for R-group decomposition, which defines the substitution sites and functional groups for each molecule. For this purpose, the MCS is mapped onto each molecule in a series and the substituents are assigned to corresponding R-groups. If there is more than one possible mapping of the MCS to a molecule, a mapping is selected for which the number of different substitution sites is minimal. For symmetrically substituted molecules, the chemical nature of substituents is used to break symmetry and identify a unique mapping of substitution sites. Figure 1 presents an example of three compounds sharing benzene as a common framework. Because of its symmetry, the framework can in principle be mapped to each molecule in six distinct ways. However, for the three molecules shown in Figure 1, only four of these six possible sites are substituted, which reduces the number of potential mappings. For molecules A and B, a consistent mapping is obtained by assigning identical substituents to equivalent sites. The methoxy group colored in red and the acetyl group (green) are unambiguously assigned to substitution sites 1 and 3, respectively. The hydroxyl group present in A and B must be assigned to the same site in both molecules; hence, it is either assigned to R-group 2 or R-group 4 in A and B. Both solutions are equivalent and thus yield an unambiguous mapping for compounds A and B. In compound C, however, the chlorine substituent differs from the hydroxyl groups of the other compounds and it is not evident whether it should be assigned to substitution site 2 or 4. Thus, in this case, the chlorine substituent is arbitrarily assigned to one of these two sites. MCS identification and R-group decomposition are carried out with Pipeline Pilot.[12]

Table 1. Data Sets[a]

activity	source	no. of compds	no. of analogue series	potency range
hydroxysteroid-17β-dehydrogenase 4 inhibitors	PubChem AID 893	400	42	25 nM−40 μM
thrombin inhibitors	PubChem AID 1215	51	6	1 nM−50 μM
cytochrome P450 3a4 inhibitors	PubChem AID 884	1251	134	25 nM−40 μM
cathepsin K inhibitors	ref 15	264	37	0.01 nM−1 mM
cathepsin L inhibitors	ref 15	290	43	0.04 nM−150 μM
cathepsin S inhibitors	ref 15	296	42	0.13 nM−1 mM

[a] Data sets containing a number of analogue series were collected from PubChem or from compound selectivity sets and served as reference for score normalization (see text for details). “no. of compds” denotes the number of compounds and “no. of analogue series” the number of analog series with distinct frameworks present in a data set.

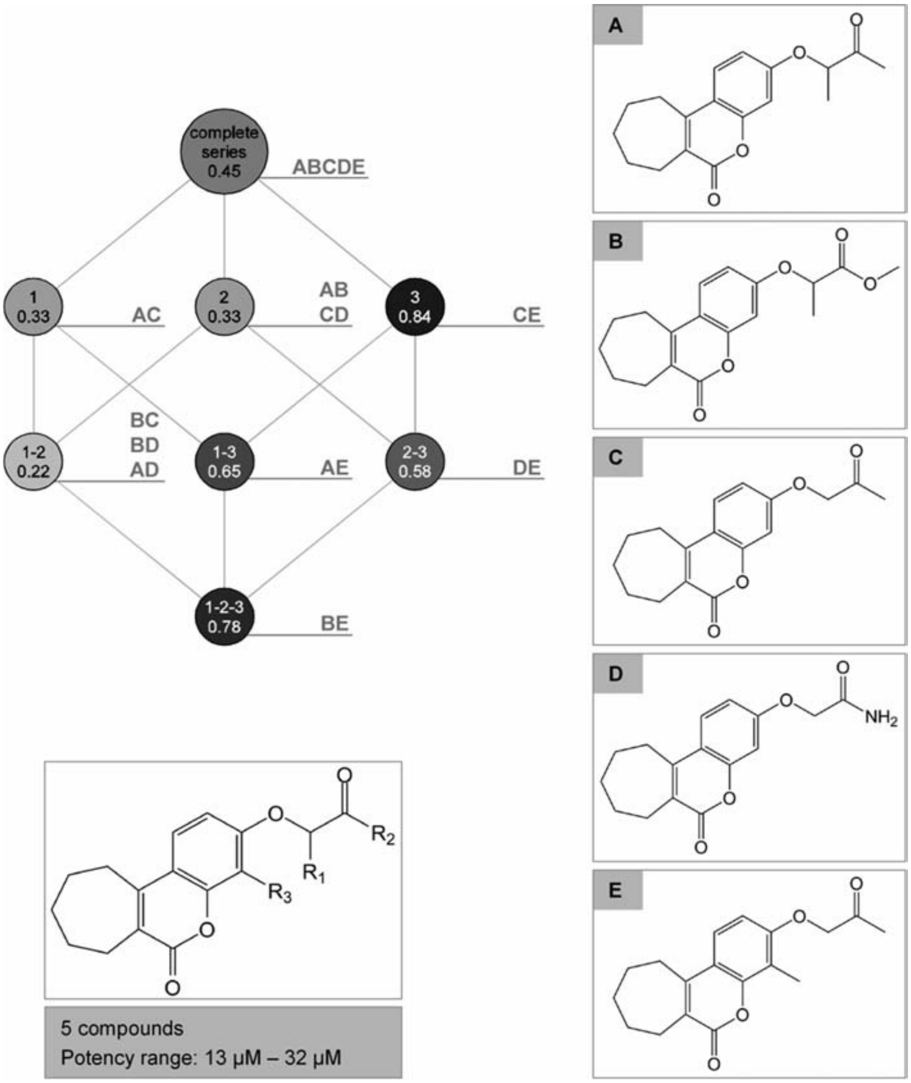


Figure 2. CAG representation of a series of hydroxysteroid-17β-dehydrogenase 4 inhibitors. Nodes correspond to compound subsets: the root node represents the entire analogue series and nonroot nodes correspond to subsets of compounds that differ only at predefined substitution sites. Node labels identify variable substitution sites and report SARI discontinuity scores calculated on the corresponding compound subsets. Nodes are color-coded according to discontinuity scores using a continuous grayscale from white for score 0 to black for score 1.

SAR Contributions from R-Groups. To assess SAR contributions of functional groups, we calculate the SARI discontinuity score for subsets of analogues that differ only at specified substituent positions. Thus, all compounds are selected from a series that have different R-groups attached to a specific site but are otherwise identical, and the discontinuity score is calculated for this subset. It follows that observed differences in SAR discontinuity can be directly attributed to R-groups at the site under consideration. Furthermore, SAR contributions from combinations of substitution sites are calculated for compounds that differ in more than one R-group position but are otherwise identical. Combinations of up to three different substitution sites are considered. For a given substitution site or combination of sites, more than one subset of compounds might exist. Discontinuity scores are then calculated for each individual compound subset and averaged to yield the final score for the substitution site combination.

The scores calculated for an analogue series are normalized to the value range [0,1], as described above, based on the distribution of subset scores within the given source data set. As summarized in Table 1, all analogue series used in this study were taken from source data sets consisting of compounds having a specific biological activity. The score distribution within such a source data set serves as the reference for score normalization of the analogue series taken from it. Accordingly, the scores are specific for a compound data set and the magnitude of scores can only be directly compared for different analogue series originating from the same set. This helps to discriminate compound series having a different degree of discontinuity and accounts for the score distribution in the entire data set. However, for analogue series taken from different data sets, the magnitude of scores cannot be compared.

Combinatorial Analogue Graphs. SAR features of analogue structures are visualized in a hierarchical CAG representation. A CAG is a graph that consists of nodes, represented by circles, and edges connecting individual nodes, drawn as lines between the circles. In general, nodes represent objects and edges a relationship between connected nodes. In a CAG, nodes correspond to compound subsets and edges indicate that compounds in connected subsets have modifications at the same substitution sites (see below). The root node represents the entire analogue series and nonroot nodes represent subsets of compounds that only differ in individual substitution sites or unique site combinations. Node labels identify the substitution sites and report discontinuity scores for the compound subset representing each site combination. Nodes are arranged in layers according to the number of substitution sites that are considered and grayscale-coded according to discontinuity scores using a continuous spectrum from white (score 0) to black (score 1). Edges are drawn from a node to all other nodes in the next layer whose substitution site combination includes all of the sites represented by the originating node. Substitution site combinations for which no compounds are available (i.e., nonexplored

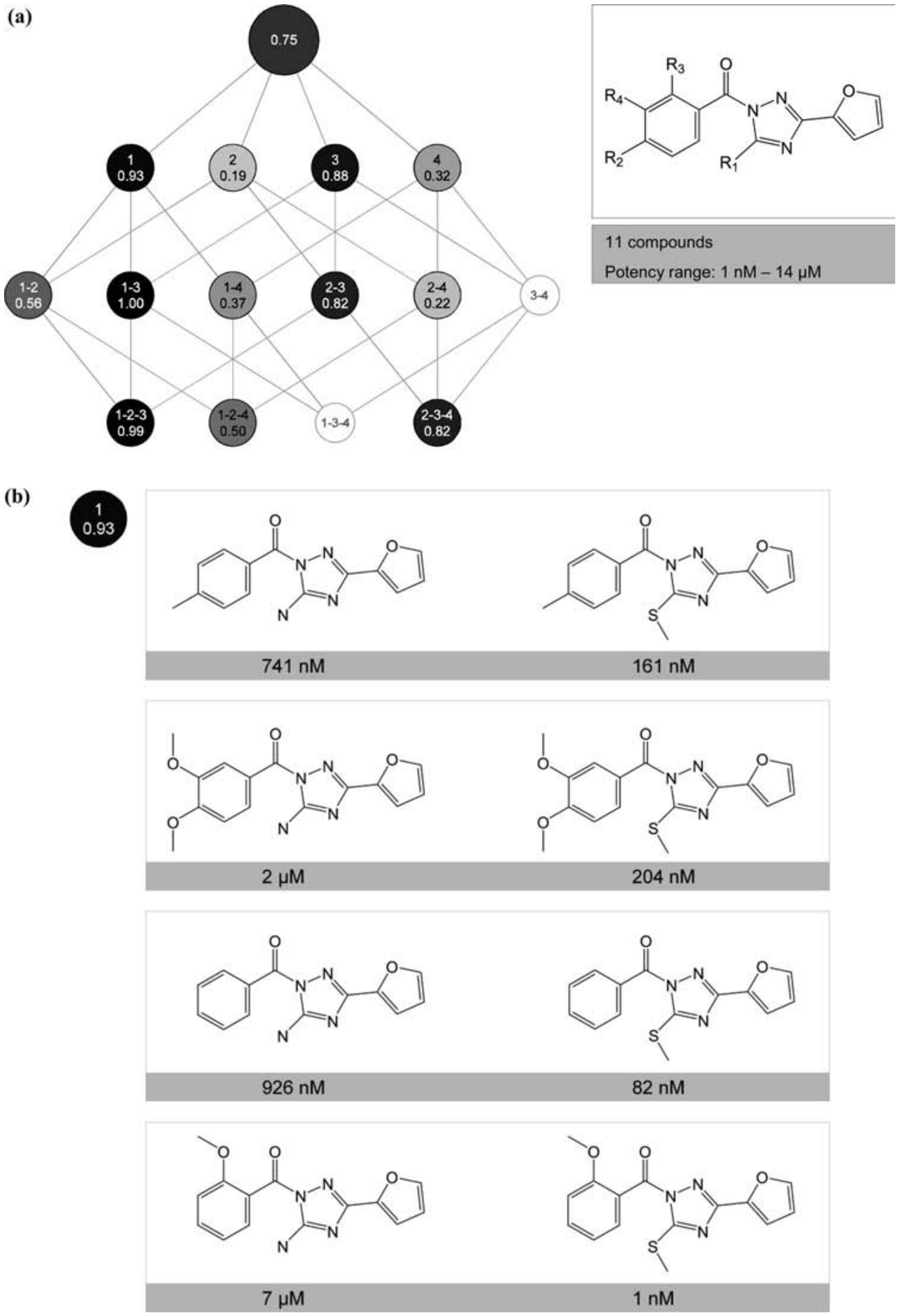


Figure 3. Thrombin inhibitors. (a) CAG representation for 11 analogous thrombin inhibitors with potency in the low nanomolar to low micromolar range. The common framework contains four substitution sites. Several nodes in the graph display considerable SAR discontinuity. (b) Pairs of compounds with modifications at substitution site 1 that form activity cliffs of increasing magnitude. The most significant activity cliff is formed by the two analogues at the bottom that have a potency difference of almost 4 orders of magnitude.

combinations) are shown as small white nodes. CAGs are calculated and displayed using R.[13] R scripts for CAG generation will be freely available via http://www.lifescienceinformatics.uni-bonn.de (see "Downloads").

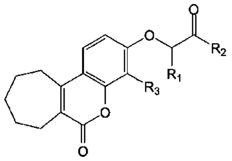
Compound Data Sets. Analogue series were extracted from screening data sets available in PubChem[14] including inhibitors of hydroxysteroid-17β-dehydrogenase-4 (AID 893), thrombin (AID 1215), and cytochrome P450 3a4 (AID 884). Compounds considered to be inactive under screening conditions on the basis of chosen activity thresholds were utilized to define the activity baseline for our analysis. In addition to analogue series collected from screening data, inhibitors of cathepsin K, L, and S were taken from compound sets designed for chemical biology applications that included highly selective and optimized compounds.[15] Table 1 summarizes the data sets used in this study, and Table 2 presents the SAR data for the analogue series that were analyzed.

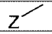
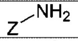
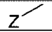
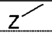
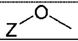
Results and Discussion

A primary goal of our analysis has been to systematically evaluate the SAR contributions of combinatorial R-group patterns in analogue series and identify substitution sites that are SAR determinants and preferred targets for further chemical exploration. The CAG-SARI approach presented herein com-

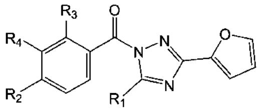
Table 2. SAR Data[a]

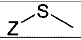
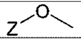
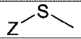
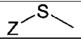
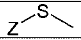
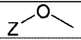
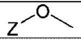
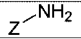
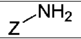
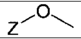
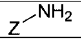
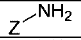
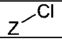
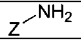
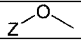
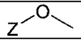
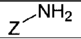
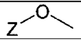
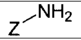
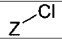
(a) Hydroxysteroid-17β-dehydrogenase 4 inhibitors (PubChem AID 893)



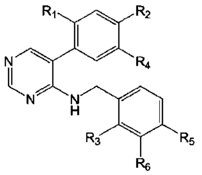
PubChem CID	R1	R2	R3	Potency [μM]
662549				32
890163				25
890639				13
2938438				25
2938604				32

(b) Thrombin inhibitors (PubChem AID 1215)



PubChem CID	R1	R2	R3	R4	Potency [nM]
977140					1
1088427					82
1088428					159
976363					204
828590					741
1084416					828
828588					926
828591					1462
969825					2227
969710					6933
828593					13951

(c) Cytochrome P450 3a4 inhibitors (PubChem AID 884)
(i)



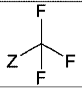
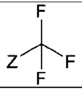
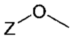
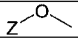
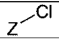
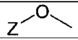
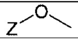
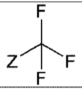
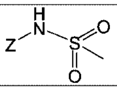
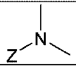
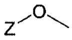
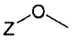
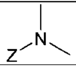
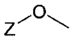
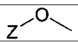
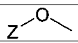
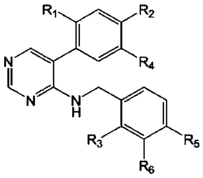
PubChem CID	R1	R2	R3	R4	R5	R6	Potency [nM]
3235235							79
3235476							79
3234995							100
3234666							126
3235489							126
3234829							158
3232982							199
3233999							199
3234568							199

Table 2. Continued[a]

(c) Cytochrome P450 3a4 inhibitors (PubChem AID 884)

(i)



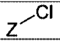
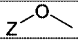
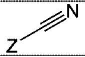
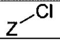
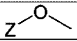
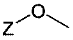
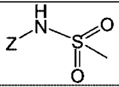
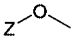
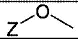
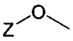
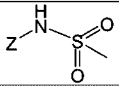
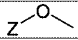
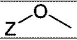
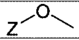
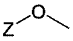
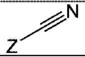
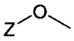
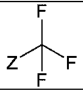
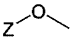
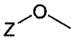
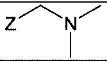
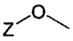
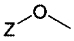
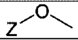
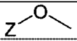
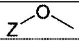
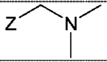
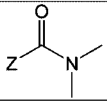
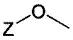
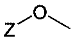
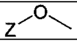
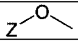
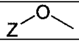
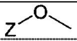
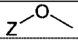
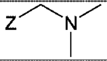
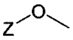
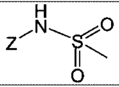
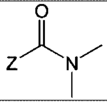
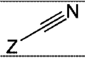
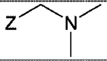
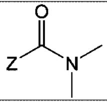
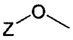
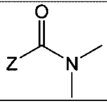
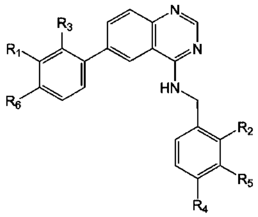
PubChem CID	R1	R2	R3	R4	R5	R6	Potency [nM]
3234784							199
3234813							199
3235150							199
3232886							251
3233050							251
3235328							251
3232698							316
3233287							316
3233374							316
3234593							316
3235521							316
3233147							398
3233799							398
3233983							398
3234079							398
3235193							398
3233488							501
3234501							501
3235200							501
3232915							631
3233258							631
3235026							631
3234812							1259
3234434							1995
3232748							2512
3234651							2512

Table 2. Continued[a]

(c) Cytochrome P450 3a4 inhibitors (PubChem AID 884)

(ii)



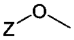
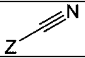
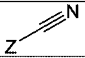
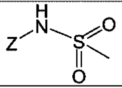
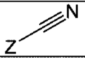
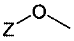
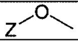
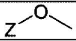
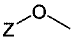
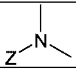
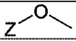
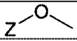
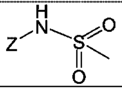
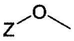
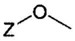
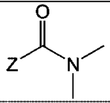
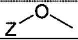
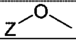
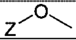
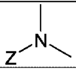
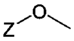
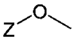
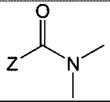
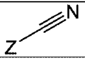
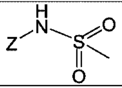
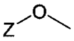
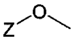
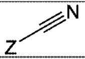
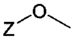
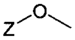
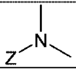
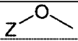
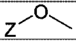
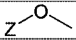
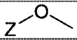
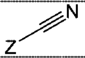
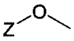
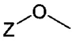
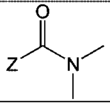
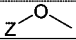
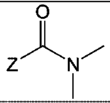
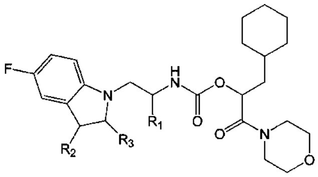
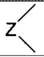
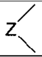
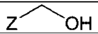
PubChem CID	R1	R2	R3	R4	R5	R6	Potency [μM]
3234444							1
3234994							1.2
3233006							1.6
3234033							1.6
3234210							1.6
3234087							2
3234476							2
3235385							2
3235533							2
3232936							2.5
3233864							2.5
3234498							2.5
3234623							2.5
3232763							3.2
3234222							3.2
3234265							4
3235512							4
3233879							5
3235481							6.3
3232832							8
3233507							8
3235397							10
3232806							12.6
3235135							12.6
3233007							15.8
3234648							15.8

Table 2. Continued[a]

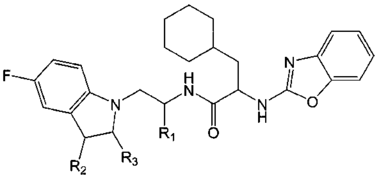
(d) Cathepsin inhibitors

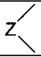
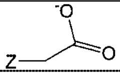
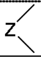
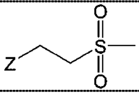
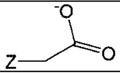
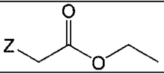
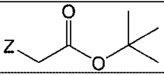
(i)



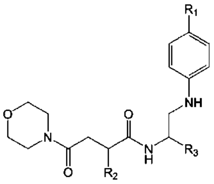
Compound	R1	R2	R3	Potency cat K [nM]	Potency cat L [nM]	Potency cat S [nM]
1				100000	100000	19
2				100000	100000	80
3				30000	100000	143
4				30000	30000	226
5				30000	30000	2950

(ii)



Compound	R1	R2	R3	Potency cat K [nM]	Potency cat L [nM]	Potency cat S [nM]
1				3710	123	3
2				20000	10000	12
3				100000	4670	15
4				14700	849	15
5				4870	369	21
6				10000	2830	26
7				100000	70000	27
8				100000	50000	71
9				30000	9670	151
10				30000	10000	222
11				100000	100000	730
12			Z=O	100000	30000	12300

(iii)



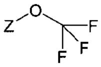
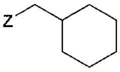
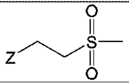
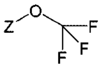
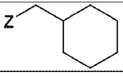
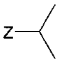
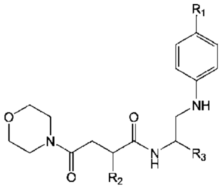
Compound	R1	R2	R3	Potency cat K [nM]	Potency cat L [nM]	Potency cat S [nM]
1				260	98	5
2				995	193	11

Table 2. Continued[a]

(d) Cathepsin inhibitors

(iii)



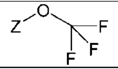
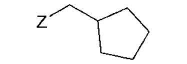
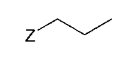
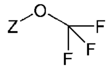
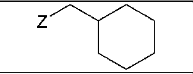
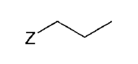
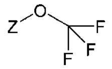
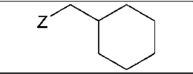
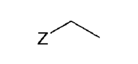
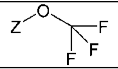
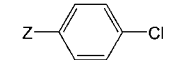
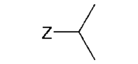
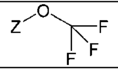
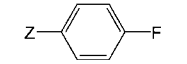
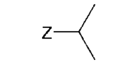
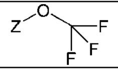
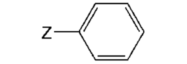
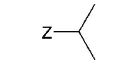
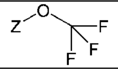
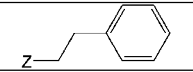
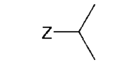
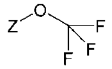
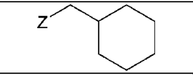
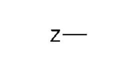
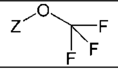
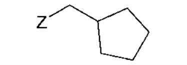
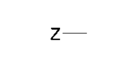
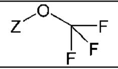
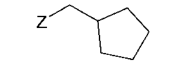
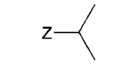
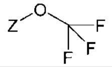
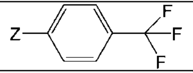
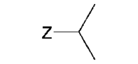
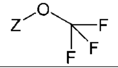
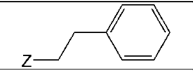
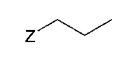
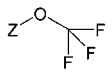
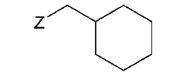
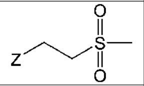
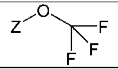
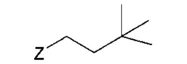
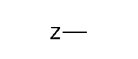
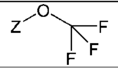
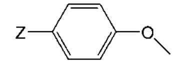
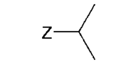
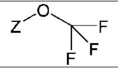
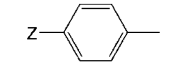
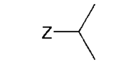
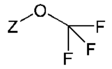
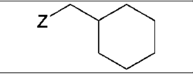
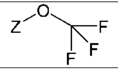
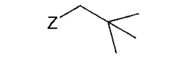
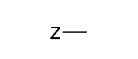
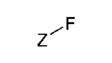
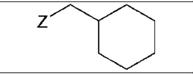
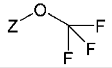
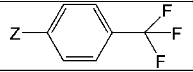
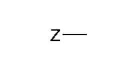
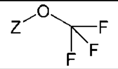
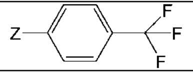
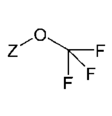
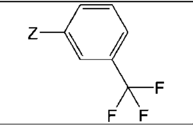
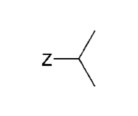
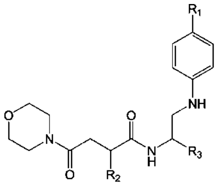
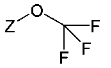
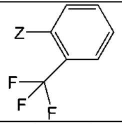
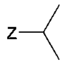
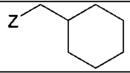
Compound	R1	R2	R3	Potency cat K [nM]	Potency cat L [nM]	Potency cat S [nM]
3				217	2678	12
4				557	285	13
5				907	267	16
6				100000	100000	21
7				11530	11530	24
8				11530	11530	25
9				84	372	27
10				24290	18900	28
11				3706	30000	31
12				1140	8990	38
13				100000	100000	43
14				291	588	60
15				5410	1520	69
16				100000	100000	70
17				30000	65000	105
18				30000	100000	134
19				100000	30000	194
20				2331	100000	408
21				30000	30000	1368
22				100000	100000	1590
23				100000	100000	15600
24				100000	30000	18530

Table 2. Continued[a]

(d) Cathepsin inhibitors
(iii)



Compound	R1	R2	R3	Potency cat K [nM]	Potency cat L [nM]	Potency cat S [nM]
25				100000	100000	100000
26				100000	100000	100000

[a] For each analogue series discussed in the text, molecular frameworks and consistently numbered R-groups are provided. For individual compounds, substituents and potency values are reported. Compounds from PubChem bioassay data are identified by their unique PubChem CID. Compounds from selectivity data sets[15] are identified by an arbitrarily assigned index. Attachment points are marked with “Z”.

bines the hierarchical organization of analogue series according to substitution site combinations with a quantitative SAR analysis function to assess site-dependent contributions to SAR discontinuity. The identification of local SAR discontinuity is highly relevant for lead optimization because increasing SAR discontinuity is thought to be related to the probability that a compound series can be further optimized. Analogues in CAG representations only differ in specified substitution sites and thus the local discontinuity scores directly reflect SAR contributions of R-groups at these sites.

Combinatorial Analogue Graphs. To illustrate our compound organization scheme, Figure 2 shows a prototypic CAG representation generated for five exemplary hydroxysteroid-17β-dehydrogenase 4 inhibitors with three substitution sites. The root node at the top represents the entire compound set and reports its discontinuity score (a score of 0.45 reflects intermediate SAR discontinuity). Each subsequent node corresponds to a unique combination of substitution sites and reports the corresponding degree of SAR discontinuity. In Figure 2, nodes are annotated with compound pairs that only differ at a specific substitution site or combination of sites. In addition, the corresponding analogues are shown. The figure also illustrates why multiple compound subsets might exist for individual nodes. For example, the compounds in pairs (A,B) and (C,D) only differ at substitution site 2 and are thus assigned to the corresponding node. However, these two compound subsets are distinguished from each other at site 1. Thus, for each pair, the discontinuity score that is only due to variation at site 2 is separately calculated and these scores are averaged to yield the final score for analogues that only differ at site 2. Hence this score reflects the overall SAR discontinuity introduced by R-group variation at site 2. For the exemplary compounds shown in Figure 2, combinations of all three sites are detected and all possible nodes are populated.

Although CAG-SARI analysis of small data sets is meaningful, larger compound series provide more SAR information for CAG representations. Because substitution sites and combinations are treated independently and only compounds with modifications at the corresponding sites are considered, data sets of increasing size and hence increasing SAR information do not introduce “background noise”.

SAR Hotspots. Figure 3a shows the CAG representation for a series of 11 thrombin inhibitors that cover a wide potency range (1 nM to 14 μM). With a discontinuity score of 0.75, the entire series shows a considerable degree of discontinuity. Several CAG nodes can be identified that are assigned high discontinuity scores, representing compound subsets with variations at substitution site 1 (nodes 1, 1−3, and 1−2−3) and at site 3 (nodes 3, 2−3, and 2−3−4). Figure 3b presents compound pairs with variations at site 1 that form activity cliffs of increasing magnitude, with potency differences of up to 4 orders of magnitude.

For data sets with a narrow potency range, the presence of similarly pronounced activity cliffs is unlikely. However, CAGs highlight the most significant discontinuity markers within a given data set and thus reveal SAR features that are characteristic for the data set. Figure 4a presents a CAG representation for 35 cytochrome P450 3a4 inhibitors that span a more limited potency range. Nevertheless, at each layer in the graph, a number of different sites or site combinations are found that produce significant SAR discontinuity, for example, nodes 2, 2−3, 2−5, and 2−3−5 or nodes 2, 6, and 1−2−6. To demonstrate the significance of SAR hotspots in CAGs for analogue prediction, the analysis was repeated after removal of the most potent compounds from the series, i.e., inhibitors with higher than 200 nM potency. Figure 4b shows the CAG recalculated for the remaining 23 active compounds. As expected, the overall discontinuity decreases due to the more limited potency range. Comparison of parts a and b of Figure 4 shows that SAR hotspots at nodes 1−2−3, 1−2−5, and 1−2−6 are retained, although in Figure 4b the most potent compounds were not taken into account. However, nodes capturing the most potent compounds in Figure 4a are now empty in Figure 4b. These nodes capture modifications at sites 2, 2−3, or 2−5. If we utilize the CAG representation in Figure 4b to predict which substitution sites should be further explored, combinations involving site 2 have high priority because this site consistently contributes to SAR hotspots and has not been thoroughly explored. Thus, we focus on site combinations capturing the most potent analogues in Figure 4a. It follows that the information provided by CAGs can be utilized to identify molecular regions where changes are most likely to introduce SAR discontinuity and yield

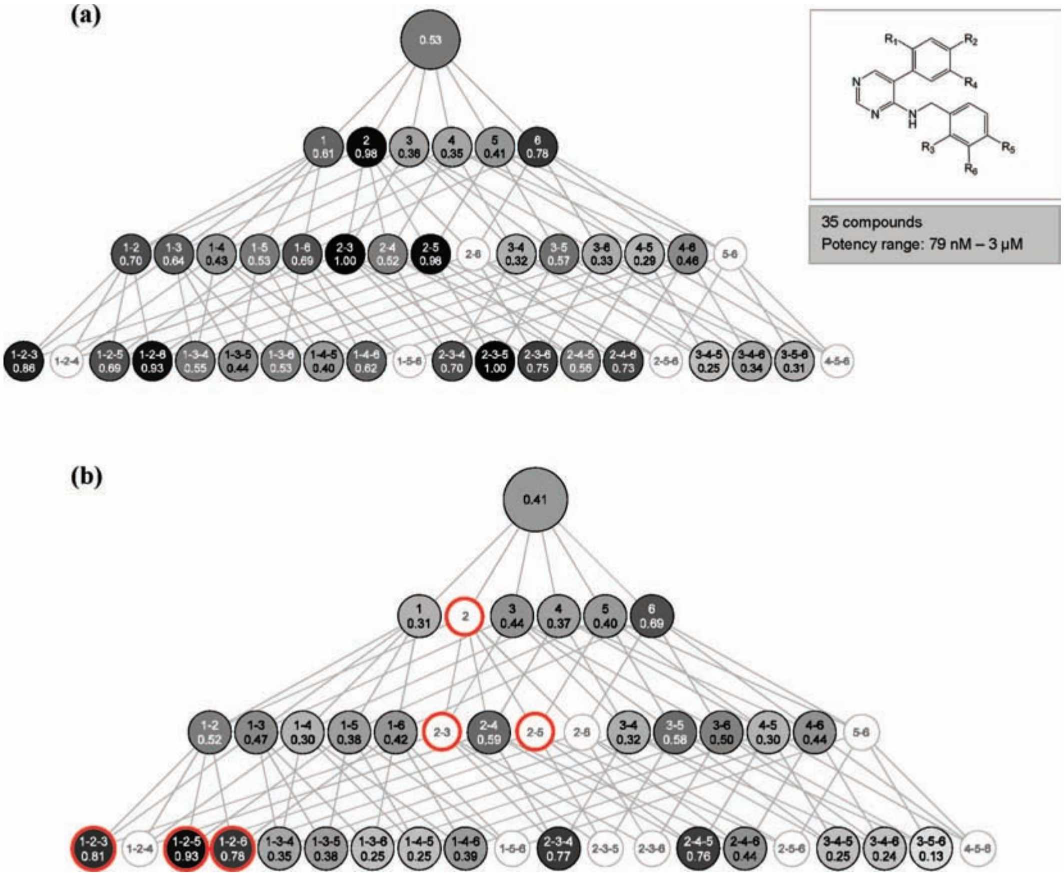


Figure 4. Cytochrome P450 3a4 inhibitors. (a) CAG representation of a series of 35 analogues. The CAG is heterogeneous and reveals large score differences between neighboring nodes. (b) CAG representation for the same analogue series after removal of 12 inhibitors with potency higher than 200 nM. Nodes discussed in the text are circled in red.

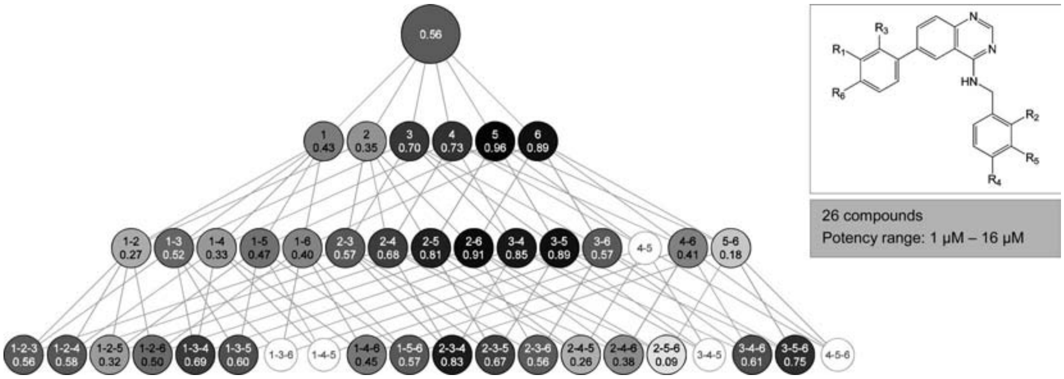


Figure 5. CAG representation for another series of inhibitors of cytochrome P450 3a4. Shown is a series of 26 analogues for which the CAG displays well-defined substitution patterns having highest discontinuity scores.

highly potent analogues. Although it cannot always be predicted precisely which substituents one should select, it is possible in many cases to combine substituents that are found at different SAR hotspots.

SAR Holes. In addition to revealing SAR hotspots, CAG analysis readily identifies SAR holes, i.e., missing substituent combinations within analogue series, as discussed above for the series represented in Figure 4. This is further illustrated in Figure 5, which also describes a set of analogues taken from the P450 3a4 screening set. The analogue series in Figure 4 and 5 contain variations at six different substitution sites. However, both series lack compounds representing variations at several substitution site pairs or triplets, which is clearly indicated in their graph representations. Thus, for practical SAR exploration, these series can be complemented with missing compound subsets in a directed manner to explore additional analogues, as discussed above. For example, in Figure 4a, substitutions at sites 2 and 6 generate considerable SAR discontinuity, but the analogue series does not contain any compound pairs with simultaneous modifications at sites 2 and 6 or 5 and 6. Moreover, in Figure 4b, site 2 and site combinations 2−3 and 2−5 (that represent the most potent compounds in Figure 4a) are SAR holes. Also, in Figure 5, site 5 represents an SAR hotspot, similar to combination 3−5, but the analogue series does not contain any compounds with 4−5 variations.

SAR Heterogeneity. Figure 4a illustrates that analogue series taken from screening data can also be rather heterogeneous in their SAR features, although the potency range is often narrow

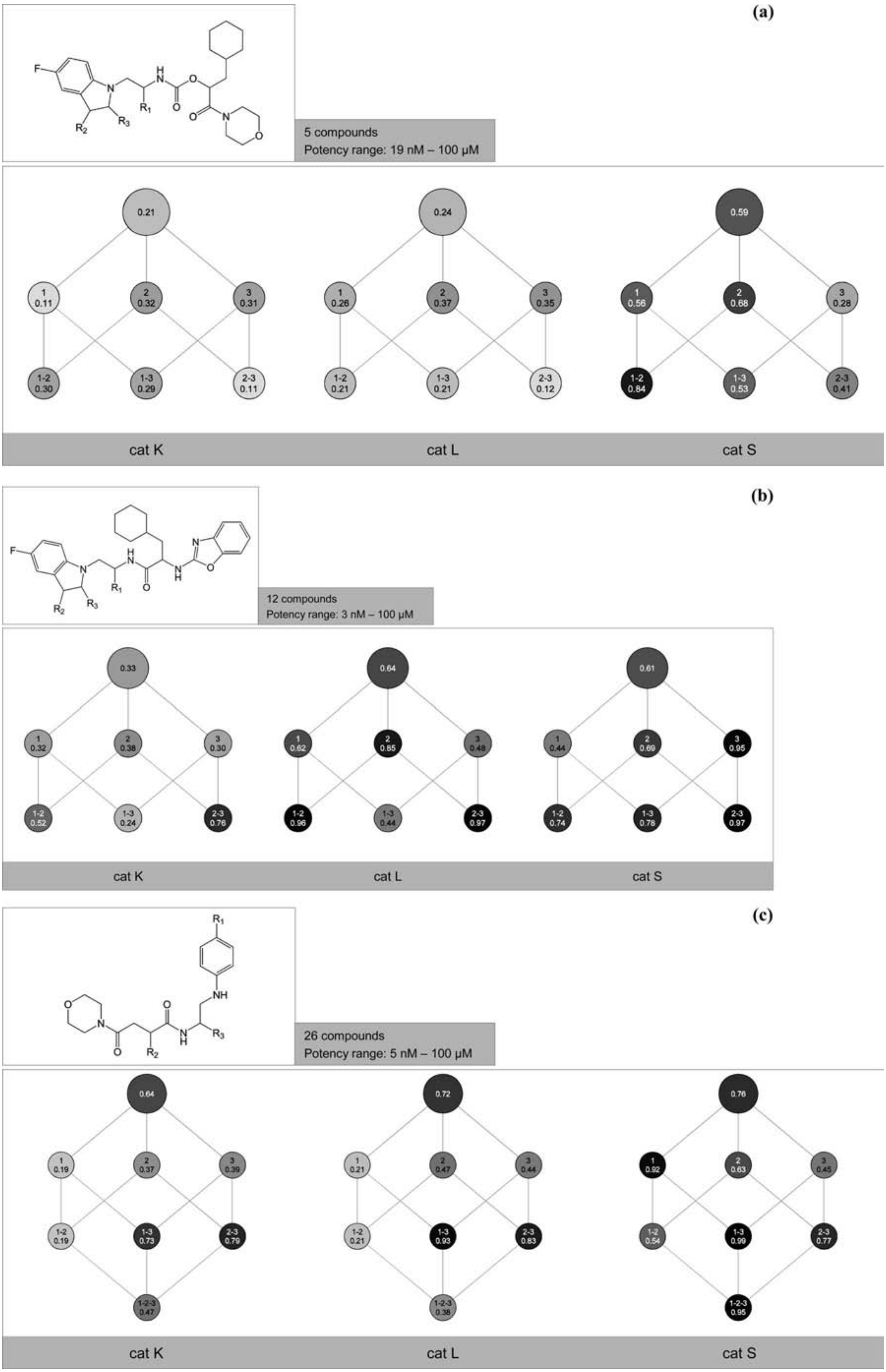


Figure 6. CAG representation for three series of compounds active against cathepsins K, L, and S. Parts (a), (b), and (c) represent graphs for individual series.

compared to lead optimization series. In the present case, the discontinuity scores of adjacent nodes vary considerably, and individual substitution sites and site combinations make significant contributions to SAR discontinuity. By contrast, in Figure 5, the distribution of node scores differs significantly from the one in Figure 4a. As shown in Figure 5, score differences between adjacent nodes are overall lower than for the alternative series. Moreover, well-defined substitution patterns exist for already available analogues that produce highest local SAR discontinuity. These patterns largely focus on sites 4 to 6 and their subsequent site combinations. Thus, comparison of the CAG representations makes it possible to prioritize analogue series on the basis of local SAR features and their distribution.

Multiple-Target SARs. In Figure 6, three different analogue series are described that were found to inhibit three related thiol proteases at significantly different levels, cathepsins (cat) K, L, and S.[15] CAG-SARI analysis makes it possible to compare such multitarget SARs. The analogue series in Figure 6a has very similar SAR characteristics for cat K and L. Scores for the entire set and all subsets are low and similar to each other. This phenotype is indicative of flat SARs that often present difficult cases for medicinal chemistry. By contrast, this series behaves differently against cat S. Here, the overall discontinuity is intermediate and there is clear SAR heterogeneity among the substitution sites and their combinations, with node 1−2 representing an apparent activity cliff. Accordingly, this series shows highest changes in potency for cat S and thus would be expected to offer greater potential for further optimization against cat S than cat K or L. Furthermore, the analogue series in Figure 6b displays similar SAR heterogeneity of substitution site combinations against cat L and S but differs in the behavior against cat K. In the latter case, SAR discontinuity and node heterogeneity is much reduced compared to the other two enzymes. Moreover, the analogue series described in Figure 6c is characterized by a significant degree of SAR discontinuity against all three enzymes, in particular, against cat L and S. For example, site 1 analogues have low discontinuity for cat L but high discontinuity for cat S, whereas site combination 1−3 represents an activity cliff in both cases, in marked contrast to site 3 analogues that cause only medium discontinuity against both enzymes. Furthermore, site combination 1−2−3 produces only a significant degree of SAR discontinuity for cat S but not cat L. Thus, for these two enzymes, the contributions of R-group combinations at different sites are nonadditive for analogues belonging to this series. Moreover, the SAR characteristics of substitution sites and site combinations 1, 3, 1−3, and 1−2−3 are found to differ significantly. Taken together, CAG-SARI analysis of compound sets that are active against multiple targets highlights substitution patterns that lead to varying degree of SAR discontinuity in different targets and identifies targets for which a given series shows the highest discontinuity and optimization potential.

Conclusions

By organizing analogue series in combinatorial analogue graphs and applying a simple and robust scoring scheme, SAR contributions of substitution sites and their combinations can be quantitatively analyzed in a systematic manner. The graph representations introduced herein make it possible to analyze the distribution of substitution site combinations in a straightforward and intuitive manner and also identify undersampled SAR regions, even if analogue series contain many compounds. Furthermore, site combinations are determined that make largest local contributions to SAR discontinuity and form activity cliffs, which are prime targets for chemical optimization efforts. Analogue series taken from screening sets we have studied contained SAR hotspots and were characterized by in part substantial SAR variability, regardless of their potency ranges. Moreover, it is possible to compare multitarget SARs including series of highly optimized compounds and describe differential SAR characteristics in detail. Taken together, our findings suggest that the CAG-SARI approach has the potential to significantly aid in extracting SAR information from different compound sources that can be utilized in hit-to-lead or lead optimization projects.

Acknowledgment. L.P. is supported by Boehringer Ingelheim.

References

(1) Esposito, E. X.; Hopfinger, A. J.; Madura, J. D. Methods for applying the quantitative structure-activity relationship paradigm. *Methods Mol. Biol.* **2004**, *275*, 131–214.

(2) Maggiora, G. M. On outliers and activity cliffs - why QSAR often disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535.

(3) Kubinyi, H. Similarity and Dissimilarity. A Medicinal Chemist's View. *Persp. Drug Discovery Des.* **1998**, *9−11*, 225–252.

(4) Peltason, L.; Bajorath, J. Molecular Similarity Analysis Uncovers Heterogeneous Structure-Activity Relationships and Variable Activity Landscapes. *Chem. Biol.* **2007**, *14*, 489–497.

(5) Eckert, H.; Bajorath, J. Molecular Similarity Analysis in Virtual Screening: Foundations, Limitations, and Novel Approaches. *Drug Discovery Today* **2007**, *12*, 225–233.

(6) Peltason, L.; Bajorath, J. SAR Index: Quantifying the Nature of Structure-Activity Relationships. *J. Med. Chem.* **2007**, *50*, 5571–5578.

(7) Guha, R.; Van Drie, J. H. Structure−Activity Landscape Index: Identifying and Quantifying Activity Cliffs. *J. Chem. Inf. Model.* **2008**, *48*, 646–658.

(8) Wawer, M.; Peltason, L.; Weskamp, N.; Teckentrup, A.; Bajorath, J. Structure−Activity Relationship Anatomy by Network-like Similarity Graphs and Local Structure-Activity Relationship Indices. *J. Med. Chem.* **2008**, *51*, 6075–6084.

(9) *MACCS Structural Keys*; Symyx Software: San Ramon, CA.

(10) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.

(11) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. *J. Med. Chem.* **1996**, *39*, 2887–2893.

(12) *SciTegic Pipeline Pilot Student Edition, version* 6.1.5; Accelrys, Inc.: San Diego, CA, 2007.

(13) R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical ComputingVienna, Austria, 2008.

(14) PubChem; http://pubchem.ncbi.nlm.nih.gov/.

(15) Stumpfe, D.; Geppert, H.; Bajorath, J. Methods for computer-aided chemical biology. Part 3: analysis of structure−selectivity relationships through single- or dual-step selectivity searching and Bayesian classification. *Chem. Biol. Drug Des.* **2008**, *71*, 518–528.

JM900107B