Similarity−Potency Trees: A Method to Search for SAR Information in Compound Data Sets and Derive SAR Rules

Mathias Wawer 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

Received May 19, 2010

An intuitive and generally applicable analysis method, termed similarity−potency tree (SPT), is introduced to mine structure−activity relationship (SAR) information in compound data sets of any source. Only compound potency values and nearest-neighbor similarity relationships are considered. Rather than analyzing a data set as a whole, in part overlapping compound neighborhoods are systematically generated and represented as SPTs. This local analysis scheme simplifies the evaluation of SAR information and SPTs of high SAR information content are easily identified. By inspecting only a limited number of compound neighborhoods, it is also straightforward to determine whether data sets contain only little or no interpretable SAR information. Interactive analysis of SPTs is facilitated by reading the trees in two directions, which makes it possible to extract SAR rules, if available, in a consistent manner. The simplicity and interpretability of the data structure and the ease of calculation are characteristic features of this approach. We apply the methodology to high-throughput screening and lead optimization data sets, compare the approach to standard clustering techniques, illustrate how SAR rules are derived, and provide some practical guidance how to best utilize the methodology. The SPT program is made freely available to the scientific community.

INTRODUCTION

Evaluating structure−activity relationships (SARs) of small molecules is a key issue in high-throughput screening (HTS) data analysis and medicinal chemistry.[1,2] SAR analysis has often different requirements. For example, in HTS data analysis, one typically needs to search for apparent and interpretable SAR information in order to prioritize hits for selection. In medicinal chemistry, one focuses on individual compound series, for example, in the course of a hit-to-lead or lead optimization project. In such cases, analogs are generated to evaluate SAR hypotheses, and computational models can be built to support the analog design process and predict the activity of test compounds.[3,4] However, even in the case of such "inductive" SAR analyses, which accumulate information focusing on one compound series at a time, it is often far from being trivial to rationalize SAR features and utilize so derived knowledge to generate highly potent compounds.

Moreover, SAR analysis becomes increasingly complicated when the focus changes from individual compound series to large compound data sets. In these cases, one is challenged with deducing SAR features from large numbers of active compounds in order to prioritize compounds for further analysis. This task applies to both HTS data and "historical" hit-to-lead or lead optimization sets where compound series and activity data have accumulated over time. Furthermore, for biological targets of high interest, hit sets from multiple screening campaigns carried out at different points in time and often using different assay formats must be analyzed in context. In such instances, a key question becomes which of typically many screening hits one should prioritize and select for further chemical exploration.

Regardless of the specific nature and origins of compound data sets under study, a general requirement for meaningful compound selection is the elucidation of interpretable SAR information, if available. Simply selecting the most potent hits as a starting point for optimization projects is generally not sufficient if SAR information is sparse or absent. Rather, compounds for which SAR characteristics are already evident in assay data are usually most likely to yield sustainable SARs and, ultimately, viable leads.

For the exploration of activity patterns and SAR features in compound data sets of any source, statistical methods,[5] graphical analysis techniques,[6,7] and compound classification approaches[8,9] have become indispensable tools. Following statistical analysis of activity data, compounds are typically clustered on the basis of structural similarity, distributions of hits over different clusters are analyzed, and representative compounds are selected. While such approaches are designed to identify structurally diverse active compounds, they are not well-suited to directly explore SAR information and deduce SAR rules. In addition to their role in HTS data analysis, graphical analysis methods are also relevant for medicinal chemistry efforts, for example, to explore substitution patterns in active compounds.[10,11] Beyond statistical analysis and compound clustering, SARs are typically explored at the structural level by focusing on maximum common substructures defining the core of analog series.[10−13]

Essentially, trying to access SAR information in compound data sets of any source requires the potency of active compounds and their chemical similarity to be systematically compared.[14,15] To these ends, SAR analysis functions have

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

10.1021/ci100197b
Published on Web 07/15/2010

Table 1. Compound Optimization and HTS Data Sets[a]

target	number of compounds	highest potency	lowest potency	average Tanimoto similarity
1. human carbonic anhydrase II	349	80 pM	33.9 μM	0.194
2. human factor Xa	874	4 pM	30 μM	0.166
3. HIV-1 protease	1243	6 fM	39 μM	0.187
4. human cytochrome P450, isoform 3A4 (AID 884)	3348	15.8 nM	39.8 μM	0.109
5. human thyroid stimulating hormone receptor (AID 938)	1765	1.3 nM	39.8 μM	0.117
6. fructose-1,6-bisphosphate aldolase [from *Giardia lamblia*] (AID 2451)	1991	44.6 nM	50 μM	0.104

[a] Inhibitor data sets were extracted from BindingDB (1−3) or PubChem (4−6).

also been introduced[14−16] that can be utilized to characterize the SAR information content of compound data sets in a global[15] or local[16] manner, analyze relationships between global and local SAR features,[17,18] characterize compound subsets,[16,19] or identify "activity cliffs" (i.e., chemically very similar compounds having large differences in potency).[20]

Organizing compound data sets according to SAR information content and/or specific SAR features goes beyond structural classification and is a prerequisite for extracting SAR information. As a first attempt in this direction, we have previously introduced a data structure termed "SAR pathways"[21] that organizes active compounds as sequences of pairwise similar molecules along potency gradients. Such SAR pathways can be systematically mined in compound data sets of any source and prioritized on the basis of fitness functions.[22] Although suitable compound series and pathways have been identified in a number of instances,[21,22] extracting SAR information through SAR pathways remains model dependent, i.e., SAR information that best fits a predefined pathway structure and its underlying parameters is prioritized.

Accordingly, we have focused on the question of how we might generalize the extraction of SAR information from compound data sets. For this purpose, several challenges must be met. If a compound data set does not contain useful SAR information, one would like to recognize this early on during the analysis. However, if it does, one would like to access available information in detail and deduce SAR rules without the need to visually inspect very many of the compounds in a data set. Furthermore, to ensure general applicability, the number of calculation parameters should be minimized and input information limited as much as possible. To these ends, we introduce a graphical analysis method termed similarity−potency Tree (SPT) that only utilizes compound potency data and similarity relationships. A similarity-based neighborhood around each compound in a data set is created using layers of nearest neighbors organized in a tree structure that mirrors local similarity and potency distributions. Overlapping compound-centric trees are systematically generated and prioritized on the basis of SAR information content. Key compound series and substitution patterns are interactively identified by analyzing individual graph representations, and intuitive SAR rules can be deduced from local neighborhoods, without the need to study the entire data set.

In the following, we first introduce the methodology and data structure and then describe how graph representations are analyzed in practice to search for and extract SAR information. By means of exemplary graphs from different inhibitor sets, including HTS and compound optimization data, we emphasize key features of the approach and demonstrate how SAR information at different levels is obtained and how SAR rules are deduced.

MATERIALS AND METHODS

Data Sets and Similarity Calculations. Activity data for inhibitors of human carbonic anhydrase II, human factor Xa, and HIV-1 protease were collected from BindingDB[23] and HTS screening data sets for human cytochrome P450 isoform 3A4, human thyroid stimulating hormone receptor, and fructose-1,6-bisphosphate aldolase (from *Giardia lamblia*) from PubChem[24] (see Table 1 for details). Only compounds with higher than 50 μM potency (K_i values for BindingDB, IC_{50} values for PubChem) were selected. If multiple K_i/IC_{50} values were reported for a single compound, their arithmetic mean was calculated. Structural similarity between compounds was assessed by calculating the Tanimoto coefficient (Tc)[25] for stereochemistry-aware extended connectivity fingerprints[26] of bond diameter four (ECFP4#S), as implemented in Pipeline Pilot.[27] We selected the extended connectivity fingerprint (ECFP) fingerprint for this study because of its high structural resolution, which helps to distinguish between closely related structures. However, dependent on the nature of compound data sets, other structural representations might also be desirable. In principle, the SPT approach can be applied using any molecular representations and similarity measures. The current implementation accepts any fingerprint representation or a precalculated similarity matrix as input.

Generation of Similarity−Potency Trees. An SPT represents the similarity and potency relationships for a subset of compounds that are structurally related to a reference compound. To select this set of compounds we define the neighborhood of a reference compound as the subset of all compounds whose similarity to the reference molecule exceeds an ECFP4#S−Tc threshold value of 0.4 (Figure 1a). One individual compound neighborhood is thus organized in one SPT data structure. To analyze a complete data set, each individual compound is chosen once as a reference molecule for the calculation of its neighborhood. Therefore, a tree is calculated for every compound in a data set. In these trees, nodes represent compounds and edges connecting the nodes reflect structural nearest neighbor relationships between the corresponding compounds. The reference compound forms the root of the tree. As such, SPTs provide a compound-centric organization of data sets and represent a previously unexplored data structure.

Two basic construction rules and two "tie-break" rules are applied for the generation of SPTs.

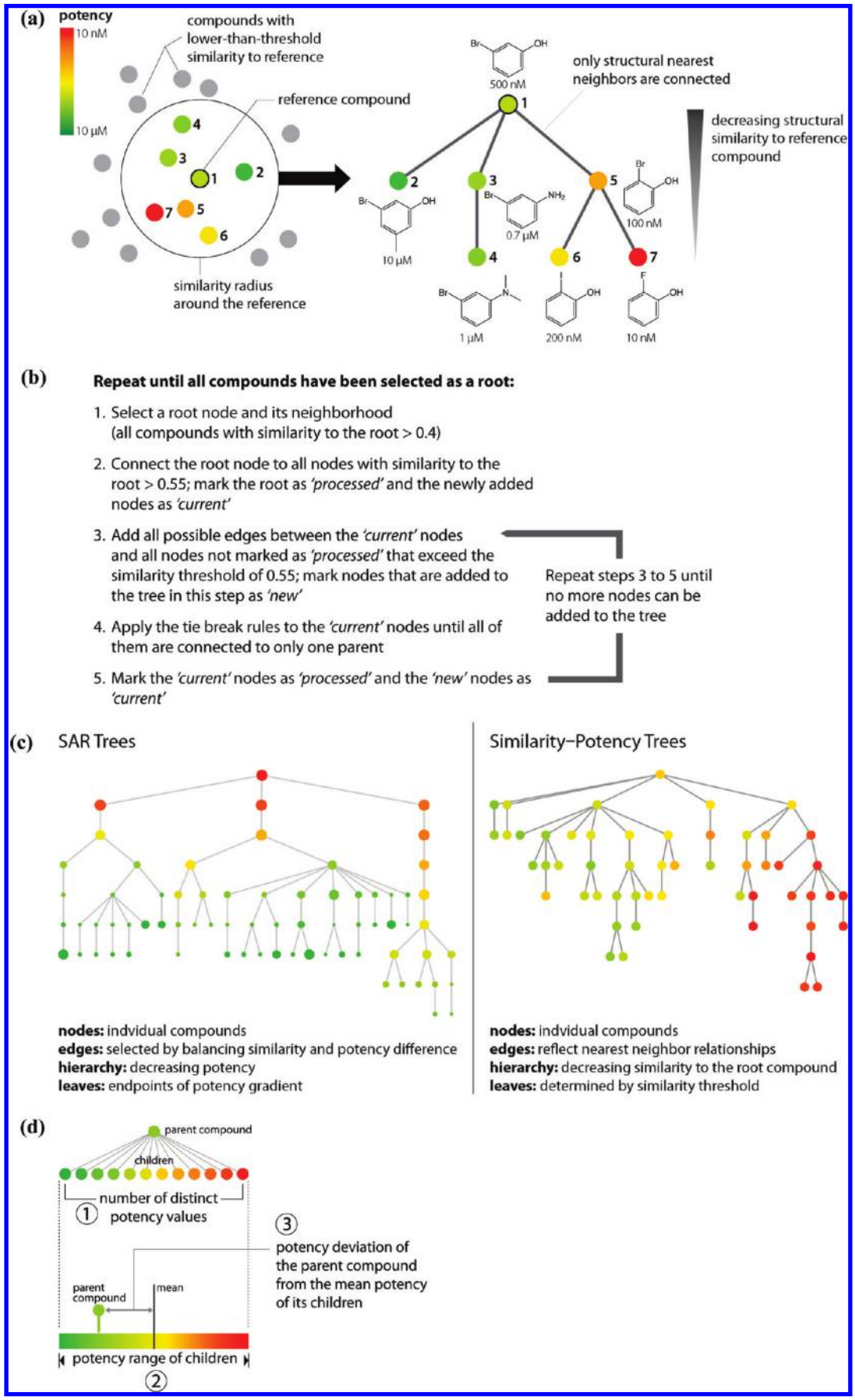

Figure 1. SPT data structure. (a) The construction of the SPT is illustrated by a schematic drawing, and a model tree is shown. On the left, a data set is indicated by nodes representing compounds. The distance between the nodes reflects their similarity. The colored nodes surrounded by the circle thus represent the compounds from the structural neighborhood of the node (compound) labeled with 1. From this set of compounds, the tree on the right is constructed. (b) A flowchart of the algorithm used for tree construction is shown. (c) The SPT data structure is compared to SAR pathways. (d) The SPT tree scoring scheme is illustrated.

1. Rule: Compounds are organized according to decreasing similarity to the root molecule along descending branches of the tree. This means that the similarity of every compound in the tree hierarchy to the root molecule must be lower than the similarity of its predecessor (unless a compound is directly connected to the root).
2. Rule: Compounds are only connected to their nearest neighbors (i.e., most similar compounds). For the examples discussed herein, compounds were only considered nearest neighbors if they exceeded an ECFP4#S−Tc threshold value of 0.55.

To generate a valid tree structure, every node (except the root) must be connected to exactly one parent. This is not ensured by the two basic construction rules alone because a compound might have equal similarity values to several other compounds at higher positions in the tree hierarchy. In this case, in order to select a unique nearest neighbor, two tie break rules are applied:

1. Tie breaker: Only the connection with the smallest potency difference between a parent and child node is retained. If this rule does not lead to the identification of a single parent (i.e., if at least two possible predecessors have the same potency), then the second tie break rule is applied:
2. Tie breaker: It is determined whether one of the potential parents is connected to the root with fewer edges than the others. If so, then only the edge to this parent node is retained. Otherwise, an arbitrary choice is made.

The algorithm used for tree construction is outlined in Figure 1b.

The similarity threshold values were chosen empirically and can be adjusted for different data sets in order to generate interpretable trees. In general, the nearest-neighbor threshold (0.55) primarily influences the speed of tree calculations. Compounds falling outside this nearest-neighbor threshold will not be part of the tree (and will be separately reported as 'singletons').

By contrast, the neighborhood threshold (0.4) determines the structural range covered by an individual tree. We used the ECFP4#S threshold of 0.4 to allow for structural variations within a tree, while retaining overall structural resemblance to the root compound.

Tree construction, annotation, and graphical analysis tools are implemented in Java using JUNG[28] and Processing[29] and enabled straightforward tree navigation. For example, tree nodes are associated with compound structures for interactive display.

Annotation of Trees. For graphical representation of SPTs, nodes are colored by the potency value of the corresponding compound. For each data set, a continuous color spectrum from green (lowest potency in a data set) to red (highest potency) is applied. Colors in trees calculated for the same data set can thus be directly compared, whereas the mapping from potency values to colors generally differs between two different data sets. In addition, all children of each node are arranged from left to right in the order of increasing potency. Numbers given below nodes designate compounds in graphical representations.

Ranking of Trees. For a data set containing k compounds, k SPTs are obtained that are ranked according to SAR information content to prioritize them for further analysis. In general, SAR information content depends on potency distributions among similar compounds. Therefore, for every node in the tree, a score is calculated based on its potency value and potency distribution of its children (nearest neighbors). The components of this score are the number of distinct potency values of the children, the potency range among the children, and the deviation of the mean potency of the children from the potency value of the parental node (Figure 1d).

Let A_i be the potency value of the compound represented by node i. For ease of notation, we define the following sets:

$$C := \{\text{children of } i\}(\text{set of children of } i) \quad (1)$$

$$P := \{A_j | j \in C\}(\text{set of potency values of the children; multiple instances of the same value are allowed}) \quad (2)$$

$$P_{\text{unique}} := \{A_j | j \in C,\ A_h \neq A_k \text{ for } h \neq k\} (\text{set of unique potency values of the children}) \quad (3)$$

Following this notation, the score for node i is calculated as

$$\text{score}_i := \frac{|P_{\text{unique}}| \times (\max(P) - \min(P))}{|A_i - \text{mean}(P)| + 1} \quad (4)$$

Thus, nodes are assigned a high score if they have multiple children with different potency values having a large spread and, in addition, a small deviation from the mean potency of their children. This means that a node is informative if it has many structural neighbors with potencies that significantly vary around its value.

The score of an entire tree (consisting of n nodes) is then simply calculated as the sum of the scores of its nodes:

$$\text{score}_{\text{tree}} := \sum_{i=1}^{n} \text{score}_i \quad (5)$$

RESULTS AND DISCUSSION

Similarity−Potency Trees. The derivation of the SPT data structure is illustrated in Figure 1a. The SPT is a graph representing the similarity and potency relationships among a set of structurally related active compounds. The compounds that form an SPT are selected from a given data set with reference to the root node. Nodes represent individual compounds and edges similarity relationships. Two nodes are connected by an edge if the corresponding compounds are nearest neighbors. Hence, the basic tree structure is determined by compound similarity relationships, and potency values are used to annotate nodes. Thus, the data structure is intrinsically simple. The length of the edges does not scale with similarity values and is only varied for graphical layout purposes. We emphasize that an SPT is generated for each compound in a data set by using it once as the root node (or reference compound). An SPT always contains the compounds from the structural neighborhood of its root compound. Thus, SPTs represent a compound-centric view of nearest-neighbor relationships. Accordingly, database compounds often occur in multiple and in part overlapping SPTs. However, compound filtering can also be applied to generate nonredundant or nonoverlapping SPTs

for comparison that focus on different compound subsets. Nonredundant SPTs are defined to permit a certain percentage of compound overlap. When moving from the root of a tree toward its leaves, structural similarity to the root node gradually decreases. Hence, the root compound provides a reference point for structural distance measurements within the tree structure. The chosen similarity threshold value only determines the boundary of the neighborhood. Because SAR information is typically sparse at low similarity values, the choice of the low similarity threshold is in general not critical for SAR analysis. Potency information is utilized to annotate SPT nodes. As expected, when scrambling experimental potency values, SPTs generally do not display interpretable SAR information (Supporting Information, Figure S1).

In Figure 1c, we compare SPTs to SAR trees that are based on a predefined SAR model.[21,22] Although the representation and layout of SPTs and SAR trees are similar, these data structures are completely distinct. In SAR trees, compound potency decreases from the root to the leaves that are the end point of a potency gradient. By contrast, in SPTs, similarity decreases from the root to the leaves that is determined by the predefined threshold value defining the structural neighborhood of the root compound. SPTs are not comparable to dendrograms generated by clustering algorithms. In dendrograms, nodes represent sets of compounds that are split into subsets when moving along the graph. By contrast, a node in an SPT represents a single compound. Furthermore, it should also be noted that the SPT concept is not a tree-based compound classification approach. It is distinct from recursive partitioning methods and decision trees[30] that classify compounds according to feature combinations and group them into terminal nodes that are determined by multiple feature decisions (or descriptor pathways). By contrast, SPT does not produce a classification of compounds into feature-dependent subsets. Rather, the SPT approach simply organizes nearest-neighbor relationships in a data set and utilizes potency as an annotation. Thus, it provides an access to similarity−potency patterns that might be contained in different data sets.

Interpretation of Trees. A prototypic SPT for a factor Xa inhibitor data set is shown in Figure 2a. Nodes at different levels in the tree gradually depart from the structure of the root. Different nodes connected to the same ancestor correspond to a subset of compounds that are nearest neighbors of this ancestor and usually derived from its structure. Given the inherently recursive nature of the SPT design, these properties give rise to two basic "reading directions" that are applied when interpreting the tree: horizontal and vertical.

The horizontal view, i.e., the comparison of compounds sharing the same ancestor, explores structural modifications of the ancestor in (most often) a series of analogs. Associated changes in potency can be directly monitored by comparing the node colors. An example is given in Figure 2b. Factor Xa inhibitor series A consists of analogs that vary at the same position, i.e., one of the nitrogen atoms of a pyrazole ring. With the exception of compounds 4 and 5, these analogs are substituted with benzene derivatives. The colors of the nodes reveal that these analogs span a wide potency range (of about six orders of magnitude) and that potency changes systematically within this series. Following the horizontal reading direction thus provides information on the number of close analogs for the respective parent compound as well as the potency distribution among these compounds. By comparing their structures, detailed SAR information can be extracted from the tree.

Children of a node might also be connected to a number of successors, which is analyzed through vertical reading following a path from the root to its leaves. This makes it possible to determine how gradual structural departures from the root affect potency. An example is shown in Figure 2c, where a path is traced. The first compound of this path (18) is derived from the root compound 1 through the addition of a fluorine atom and the replacement of a methyl group by a trifluoro-methyl group (see Figure 2a). These substitutions are retained in all of the following compounds in the path, and subsequent modifications exclusively occur in the outer benzene ring (compounds 18−24 in Figure 2c). Along the path, a gradual structural transition from an aminosulfonyl group to an *N*-methyl-piperidine ring is observed that does not alter compound potency in a significant way.

For comprehensive SAR analysis, horizontal and vertical views must be combined. This makes it possible to identify horizontally arranged "sibling" nodes, as illustrated in Figure 2b, which undergo further structural modifications along vertical paths. For example, a key feature of the SPT in Figure 2a is its separation into two subtrees representing different SAR phenotypes. The root node has two nearest neighbors with a large potency difference that are the roots of these subtrees. A striking difference between the subtrees is the distribution of potency values among their nodes. The subtree on the left contains an analog series that covers a wide potency range. By contrast, the subtree on the right almost exclusively consists of highly potent inhibitors. What are the structural origins of this potency distribution? A hypothesis can immediately be formulated by comparing the SPT root and its two children. In compound 18, and all of its successors, the 3-amino-benzisoxazole ring system of the root structure is retained, whereas it is replaced by different ring systems (or an acyclic moiety in one case) in compound 2 and its successors. In compound 17 and 32, the two most potent inhibitors in the left subtree, the 3-amino-benzisoxazole ring system of compound 1 is replaced by a benzamidine or 2-amino-isoquinoline moiety, respectively. The structural overlap between these groups suggests that benzamidine pharmacophore elements are important for activity. The only child of compound 32 differs from its parent by the absence of the amine substituent in the isoquinoline group, which causes a significant loss in potency. Additionally, all compounds with benzene derivatives carrying a substituent not similar to the amidine moiety are less potent (as further discussed below; see Figure 2c). Thus, the substituent at the pyrazole ring is in this case crucial for high potency, and 2-amino-isoquinoline (compound 32) and benzamidine (compound 17) moieties represent conservative and more favorable replacements of the 3-amino-benzisoxazole ring system of compound 1.

Formulating SAR Rules. In general, whenever an SPT displays horizontal and vertical tree patterns with directed potency changes, SAR information can be extracted and interpreted. Then, SAR rules can be formulated by combining the horizontal and vertical reading directions. As an example, we consider the interpretation of the factor Xa inhibitor set presented above. Vertical reading of the right subtree, labeled B in Figure 2a, and shown in detail in Figure 2c, reveals a

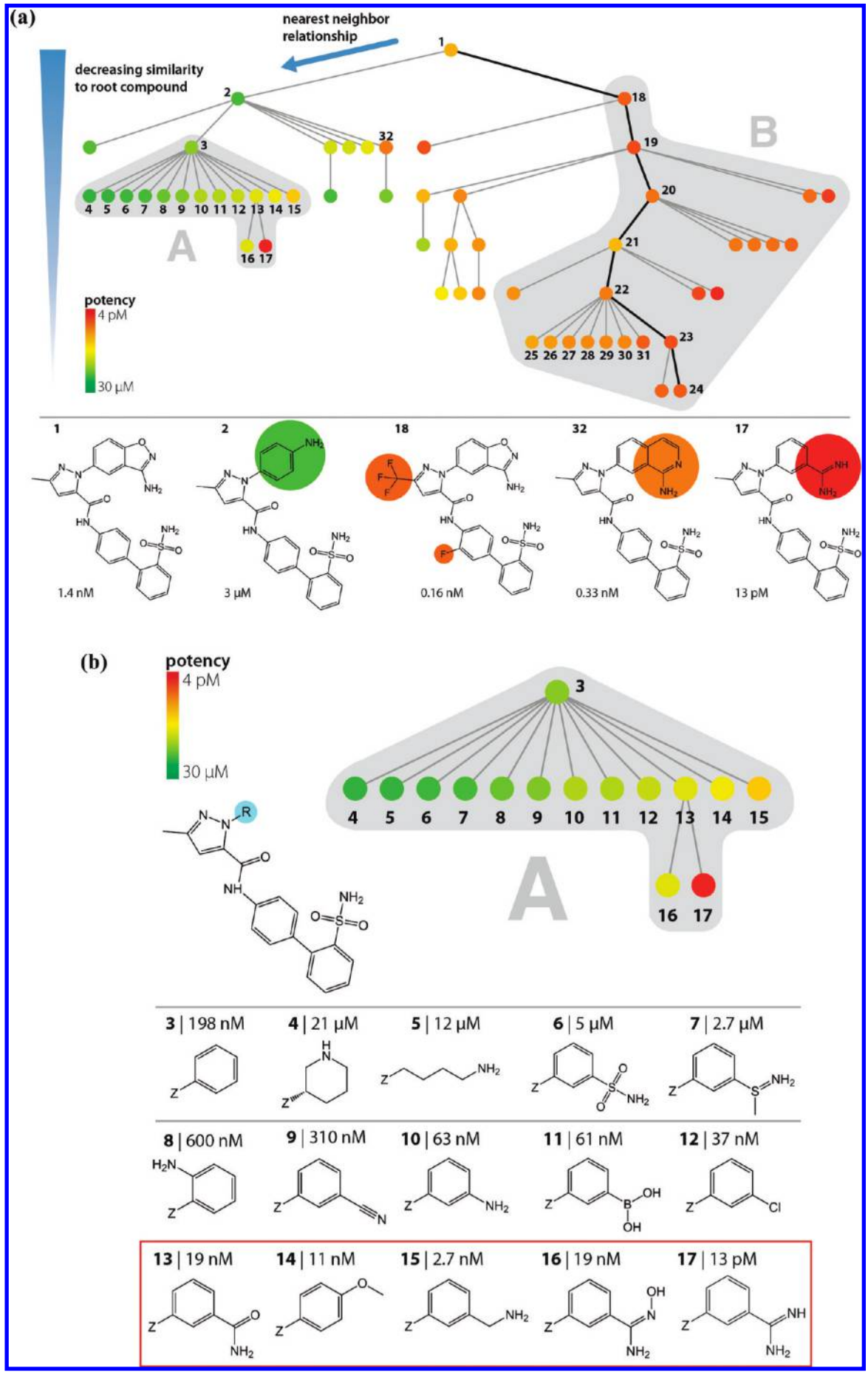
(a)
nearest neighbor relationship
decreasing similarity to root compound
potency
4 pM
30 µM
A
B
1.4 nM
3 µM
0.16 nM
0.33 nM
13 pM
(b)
potency
4 pM
30 µM
3 | 198 nM
4 | 21 µM
5 | 12 µM
6 | 5 µM
7 | 2.7 µM
8 | 600 nM
9 | 310 nM
10 | 63 nM
11 | 61 nM
12 | 37 nM
13 | 19 nM
14 | 11 nM
15 | 2.7 nM
16 | 19 nM
17 | 13 pM

Figure 2

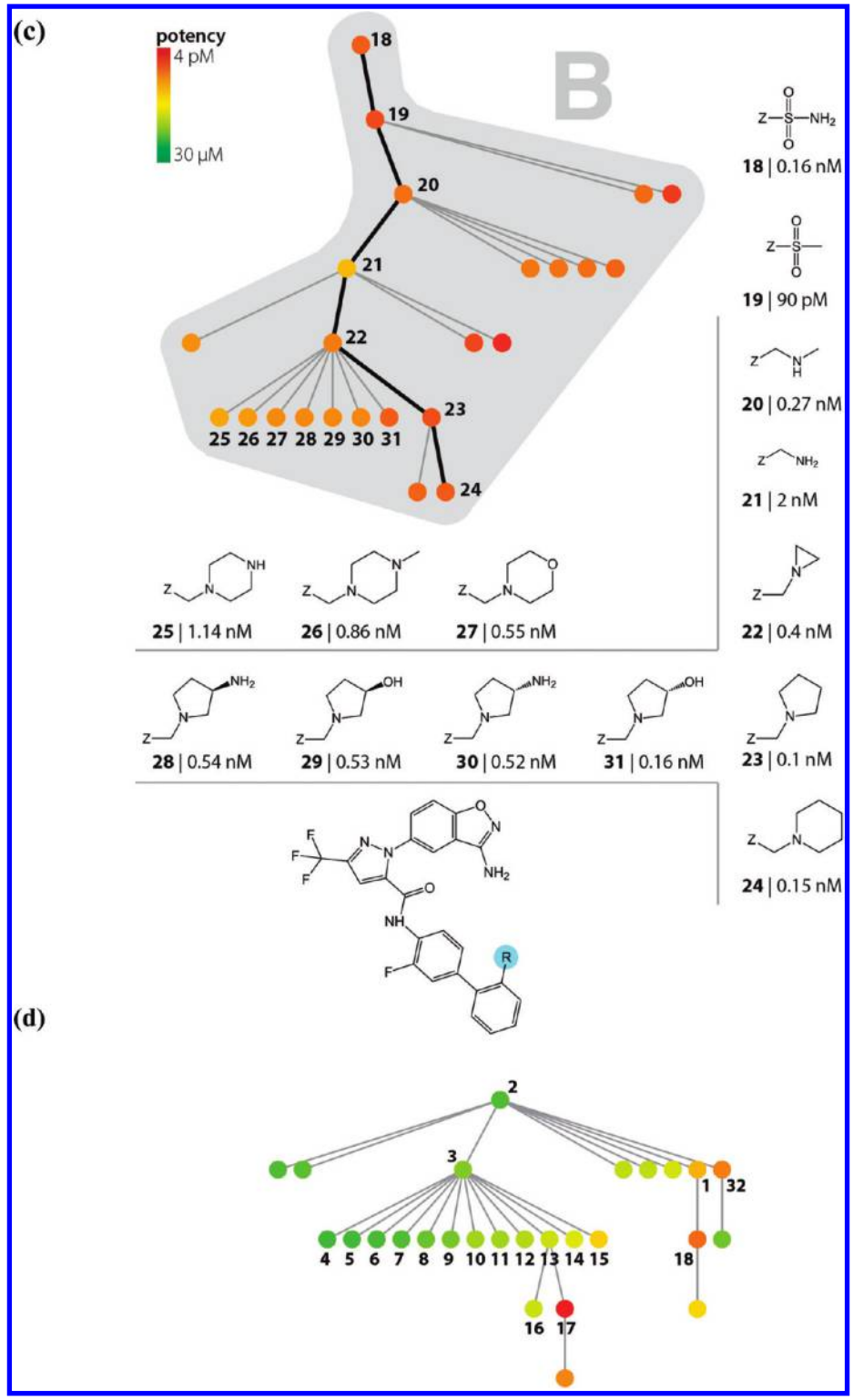

Figure 2. Prototypic SPT. An SPT for a series of factor Xa inhibitors is shown, and characteristic features are illustrated. Nodes are color coded according to compound potency using a color spectrum from green (lowest potency in the entire compound data set) to red (highest potency), as indicated by the color bars on the left. Furthermore, nodes are consistently numbered corresponding to the numbers of compounds shown in (a−c). In (a) and (c), paths from the root node to a leaf are traced with black edges, and the compounds represented by these paths are shown in (c) after R-group decomposition. In (a), subsections of the tree are highlighted that consist of two analog series. Compounds corresponding to selected tree nodes are shown, and key modifications compared to the root compound are highlighted in corresponding node colors. Series A and B are shown in (b) and (c), respectively, following R-group decomposition. Substitution sites (R) are highlighted in blue. In (b), the five most potent compounds of the series are boxed in red. All of these compounds contain substituents resembling benzisoxazole. In (d), an alternative SPT is shown that is rooted at compound 2. Potency values are reported for all compounds.

potent analog series that spans multiple nearest-neighbor relationships. The organization of nodes clearly visualizes the structural relationships among compounds in this subtree, which simplifies the analysis of structural features and potency distribution of this analog series. Nevertheless, in this particular case, the color code indicates a low variation of potency values among the compounds, which suggests that an in-depth analysis of this series alone provides little

additional SAR information. One would thus focus on the structural feature(s) that distinguishes this series from other compounds in the tree. As mentioned above, all of these analogs are distinguished from the root compound by the presence of a fluorine substituent at the central benzene moiety and a trifluoro-methyl substituent at the pyrazole ring. Hence we can conclude that the addition of fluorine substituents to the pyrazole and benzene rings is favorable. Furthermore, following the subsequent nearest-neighbor relationships down the subtree reveals a gradual transition from small sulfonamide to larger piperidine substituents (Figure 2c) with only small to moderate effects on potency indicating that potency is only little affected by substitutions at position 2 of the peripheral benzene ring, although there appears to be a slight preference for larger substituents. Horizontal reading of the left subtree, labeled A in Figure 2a and shown in detail in Figure 2b, reveals that all analogs have only low to medium potency, with two exceptions, compound 15 and in particular compound 17, both of which contain pyrazole substituents reminiscent of benzisoxazole. In fact, all of the more potent analogs in this subtree contain pyrazole substituents that show pharmacophore resemblance to benzisoxazole (highlighted by a red box in Figure 2b). Thus, we conclude that both the N-substituent at the pyrazole is important for potency and the aromatic or heterocyclic substituents should best be bioisosteric replacements of benzisoxazole. This example shows how SPT interpretation following horizontal and vertical patterns can be transformed into simple rules governing the SAR behavior of compound subsets. The combination of such rules can lead to suggestions for further analog design. For example, in this case, one would probably like to investigate whether combined fluorine substituents and benzisoxazole replacements at the pyrazole ring might further increase potency of this factor Xa inhibitor series.

SAR Information Content and Tree Ranking. In general, an SPT that is rich in SAR information is characterized by the presence of multiple nodes with children having varying potency. For a given data set, exhaustively generated SPTs are ranked for efficient analysis using a scoring function that prioritizes informative trees (see Materials and Methods Section for details). The tree score is calculated as the sum of node scores such that SPTs are prioritized that contain a large number of informative nodes, i.e., nodes that represent compounds having many analogs that display varying potency. The ranking procedure, as illustrated in Figure 1d, is sensitive to both information-rich horizontal and vertical tree arrangements. A tree node obtains a high score if: (*i*) it has many children with different potency values, (*ii*) the children span a large potency range, and (*iii*) the deviation of its potency value from the average potency value of the children is small. Thus, because tree scores are obtained by the sum of their node scores, a tree achieves a high score if it contains many high-scoring nodes. For example, the SPT shown in Figure 2a is the top-ranked tree for the factor Xa data set. By contrast, the SPT in Figure 2d that is rooted on compound 2 is ranked at position 132. This SPT essentially consists of the left subtree of the top-ranked SPT and hence contains much less information.

Highly ranked SPTs can be readily analyzed for attractive SAR patterns. In practice, it is straightforward to scan through preferred SPTs and select a subset for further analysis. If top-ranked SPTs do not display obvious SAR patterns, then it can be concluded that the data set does not contain interpretable SAR information, and the analysis can be terminated. By contrast, if SAR information is present, top-ranked SPTs provide clear indications. However, it has to be considered that SPTs are a descriptive data analysis technique and thus reflect the nature and composition of the underlying data set including possible bias. For example, in a highly skewed data set, very small compound sets that reflect a useful SAR might be difficult to identify. Due to the focus on local SARs and the relative ranking of trees, SPTs containing the informative series should nevertheless obtain a (relatively) high rank.

As discussed in the following, the analysis of alternative tree representations often reveals complementary SAR information.

Compound-Centric Analysis. Comparison of the trees in Figure 2a and 2d also illustrates the relevance of the compound-centric analysis scheme and the exploration of alternative tree representations. Depending on the root compound, the composition and information content of SPTs might substantially vary. For example, nearest neighbors, such as compounds 1 and 2 in Figure 2, might produce data structures from which either much or only little SAR information can be deduced. However, the importance of the compound-centric approach for graphical analysis goes beyond varying SPT composition. This is demonstrated in Figure 3 that shows two SPTs for HIV protease inhibitors having almost identical compound composition. In addition, both root compounds also have very similar potency. The SPT in Figure 3a displays a well-ordered potency distribution. By contrast, the potency distribution is much less systematic in the alternative tree in Figure 3b. The root and representative compounds occurring in both SPTs are shown in Figure 3c. Clearly, selecting the first compound (A) as a root produces a tree that has higher SAR information content and is easier to navigate than the tree obtained by selecting the second compound (B) as a root. Hence, compound-centric SPT analysis characterizes many overlapping neighborhoods from which those that capture available SAR information in most obvious ways can be selected.

Extracting SAR Information. The analysis of different SPTs makes it also possible to extract complementary SAR information from compound data sets. Figure 4 shows the SPT of another subset of factor Xa inhibitors, ranked at position 55, which displays a broad but not random potency distribution. Potent compounds concentrate in the right part of the tree. When comparing the root compound to its only child, compound 2, it becomes apparent that removal of the benzylether group significantly increases potency. Analyzing the potency distribution of the children of compound 2 provides additional information. The presence of seven predominantly green nodes on the left and five yellow to orange nodes on the right mirrors potency increases as a consequence of structural variations. Compared to compound 2, all weakly potent analogs have substitutions at the peripheral thiophene or piperidine ring, in contrast to highly potent compounds. In this series, the most promising region for potency-increasing substitutions is the indole ring, as revealed by compounds 4−6 and their successors. In compounds 7 and 9, the isoxazole moiety is also modified compared to compound 2. The comparison of compound 7

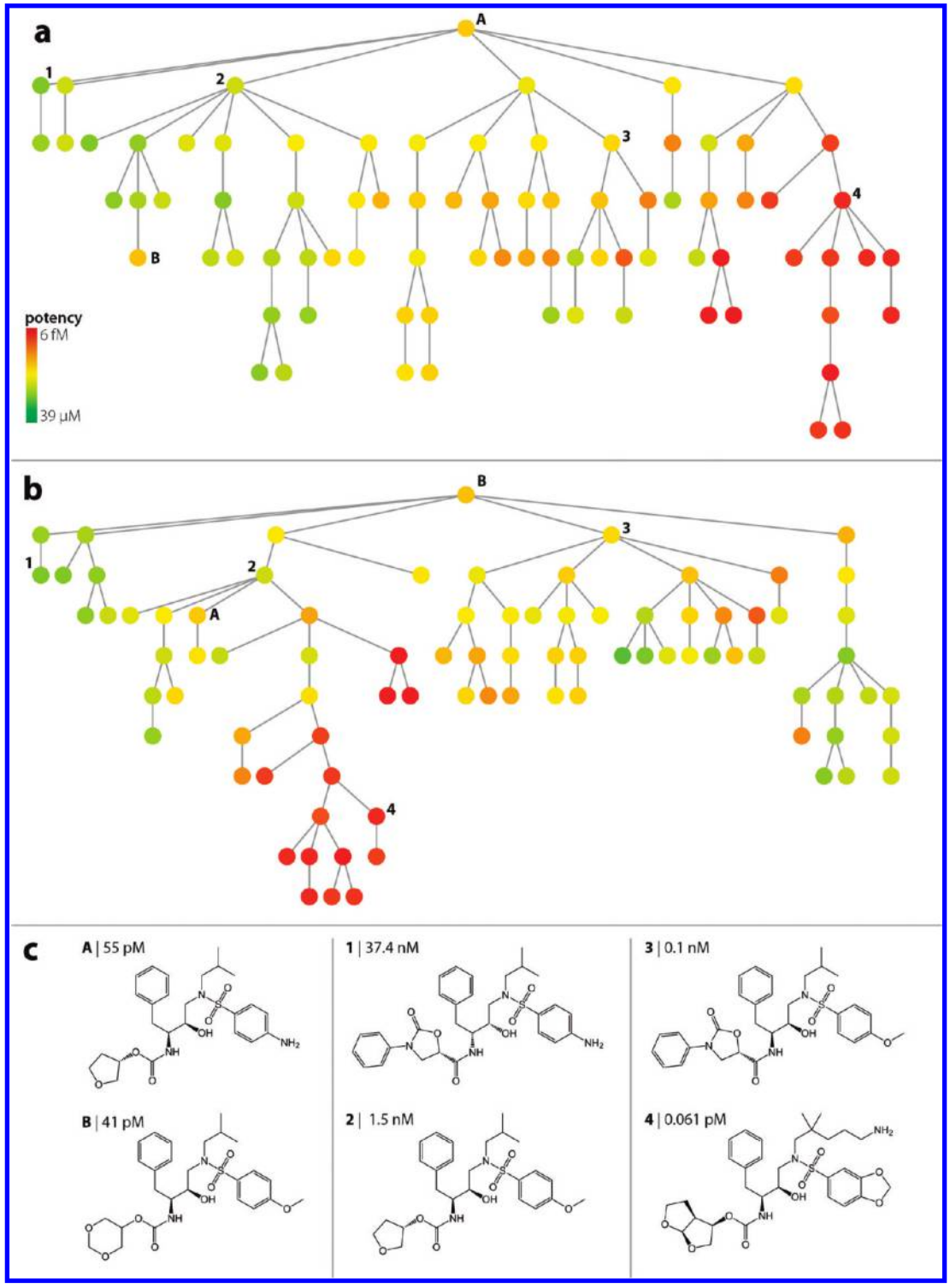

Figure 3. SPT comparison. In (a) and (b), two trees representing nearly identical subsets of HIV-1 protease inhibitors are displayed that differ in the choice of the root node. Six nodes are consistently labeled, and the corresponding structures are shown in (c). Compound potencies are reported.

and its successor (compound 8) shows that a reduction in potency is again caused by altering the thiophene ring. However, when following the path through compound 9 in Figure 4, one observes that substituting the thiophene ring does not affect—or even has a positive effect on—potency if the 1,3,4-thiadiazole ring is replaced by an amide group. Hence, this SPT is focused on different compound subsets than the top-ranked tree in Figure 2a, and detailed SAR rules can also be derived in this case, although the SPT is overall less informative than the top-ranked tree.

Figure 5 shows an SPT for carbonic anhydrase inhibitors that is only lowly ranked, for obvious reasons. In this case, the tree exclusively consists of highly potent sulfonamide derivatives. The horizontal view spanning analogs 1−6 and all vertical views provide only very little, if any SAR information. Thus, an SPT with such characteristics has low priority for analysis and is assigned a low rank.

In Figure 6, a highly ranked SPT is displayed for a subset of structurally very similar HIV-1 protease inhibitors that reveals systematic SAR characteristics. At each level of the

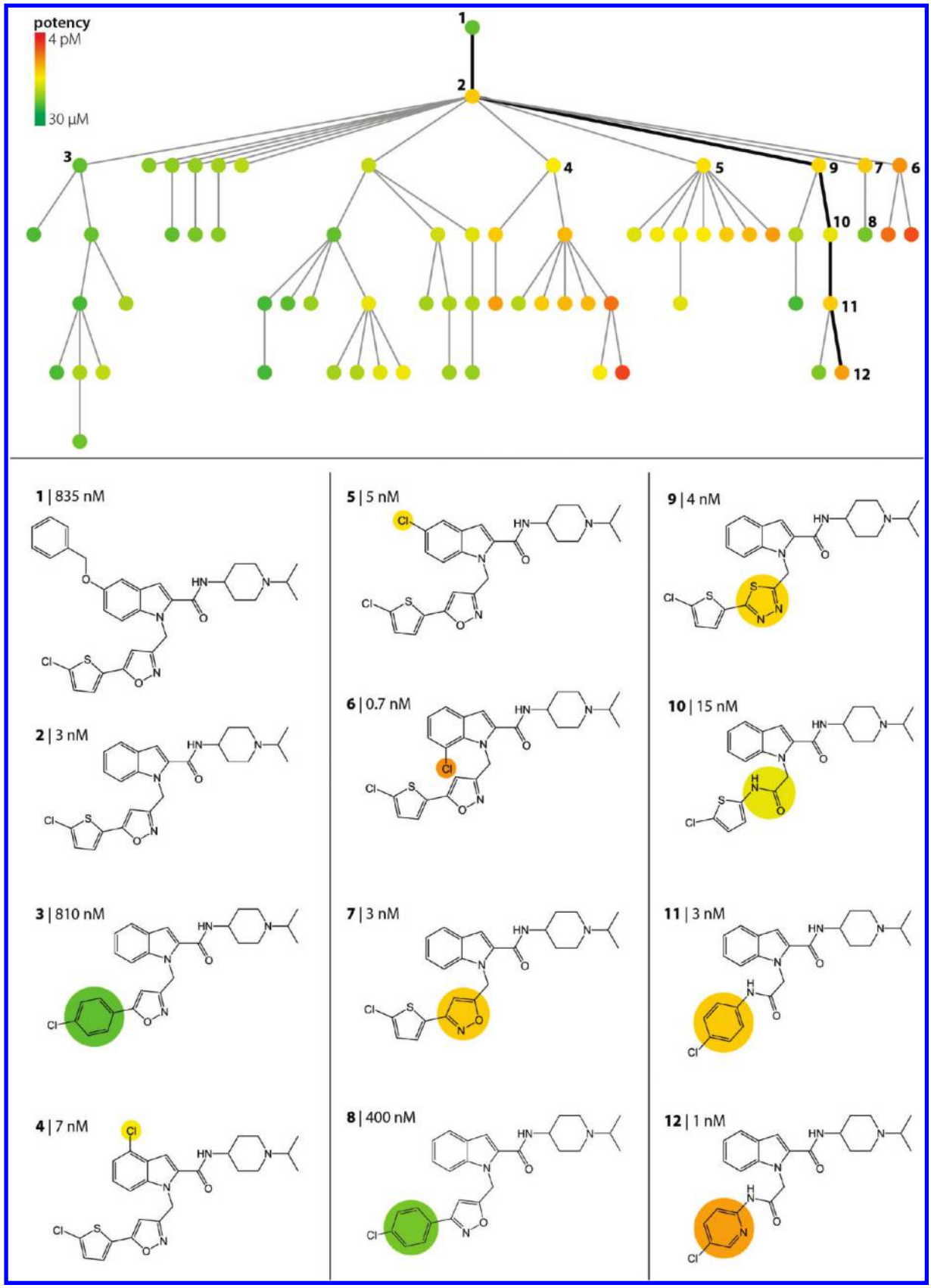

Figure 4. SPT with high SAR information content. A tree representing a subset of factor Xa inhibitors is shown. A path from the root to node 12 is traced with black edges, and structures are provided for all numbered nodes. In structures 3−12, substitutions relative to their parent compounds are highlighted in corresponding node colors. Compound potencies are reported.

tree, potency tends to increase from left to right, regardless of whether the nodes descend from the same or different parents. In addition, several pathways with compounds having either comparable or increasing potency are observed. The root compound has multiple children that form a gradient over a wide potency range (18 pM−514 nM). Most of the

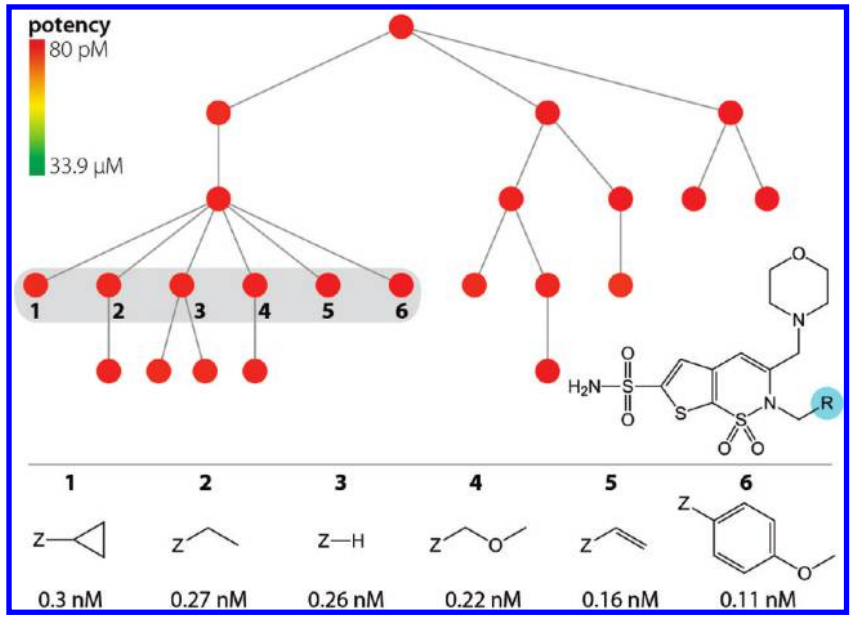

Figure 5. SPT of low SAR information content. A tree for carbonic anhydrase II inhibitors having comparably high potency is shown. A compound series is highlighted in gray, and R-groups are displayed below the common core structure. Additionally, compound potency values are reported below the R-groups.

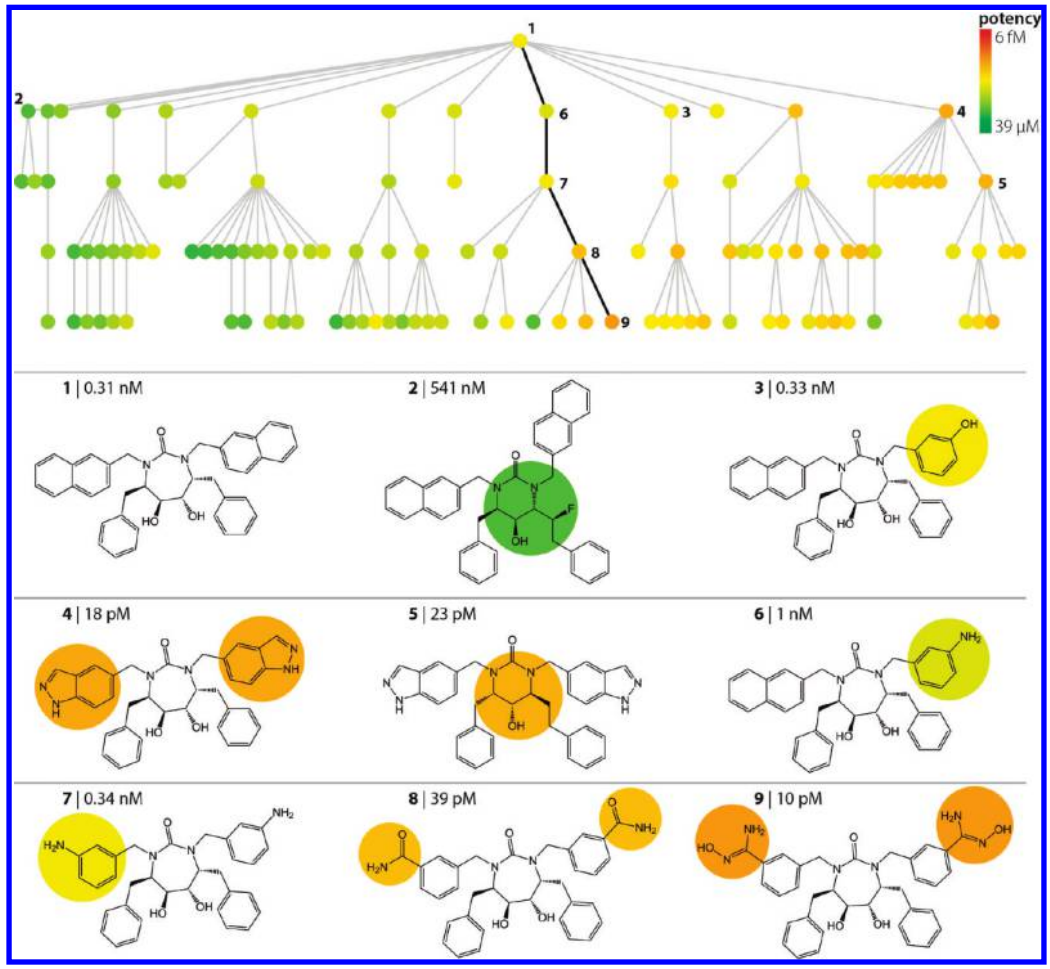

Figure 6. Well-organized SPT structure. An SPT for HIV-1 protease inhibitors is shown with many groups of sibling nodes and systematic potency distribution. A path is highlighted in black, and compound structures corresponding to numbered nodes are provided. In each compound except the root, changes relative to its parent are highlighted in corresponding node colors. Compound potencies are reported.

children also have multiple successors, resulting in a well-structured tree of balanced composition. Again, combining horizontal and vertical views provides an immediate access to SAR information. Initially, first-layer compounds and their structural differences to the root are compared. Among these, the most and least potent compounds (2 and 4, respectively)

differ in several ways. The central ring consists of six atoms in compound 2 and seven in compound 4. Because several highly potent successors of compound 4 also contain a six-membered central ring (such as, for example, compound 5), this structural difference is unlikely to account for potency changes. Therefore, substituents at the nitrogen atoms of the central ring are inspected next. Several potent analogs, but no weakly potent compounds, contain a heteroaromatic ring system or benzene derivatives (e.g., compounds 3 and 6). A path through compound 6 consists of nodes that display steadily increasing potency toward the leaf (compound 9). Evaluating structural modifications along this path reveals a critical role of two substituents. Replacing the naphthalene ring system in compound 6 with an aniline ring leads to a symmetrically substituted compound 7. Following the path to compounds 8 and 9, the two aniline amino groups further evolve into amide and hydroxy-amidine groups, respectively, and these modifications are accompanied by a steady increase in potency. Thus, following ring substitution patterns through this tree reveals interpretable SAR rules.

Analysis of HTS Data Sets. The application and analysis of SPTs discussed above for compound data sets is readily transferable to HTS data. SPTs have been calculated for a number of HTS hit sets available in PubChem, and in Figure 7 representative examples of HTS data sets are shown that have varying SAR information content. Table 1 reports the average Tanimoto similarity for all compound sets, indicating that the HTS data sets are structurally more diverse than the BindingDB compound sets, as we would expect. In Figure 7a, the SPT of the cytochrome P450 3A4 hit set reveals rich SAR information. SPTs for this data set were calculated using the same parameter settings as for the previously discussed compound data sets. Among the characteristic horizontal and vertical patterns, two series of analogs become apparent that have corresponding substitutions but two different central ring systems (one of which is a substructure of the other). Interestingly, both series essentially follow the same potency patterns, thus indicating the presence of a transferable SAR. Thus, in this case, high SAR information content of HTS data becomes immediately apparent on the basis of SPT analysis. In Figure 7b, a highly ranked SPT for thyroid stimulating hormone agonists is shown that contains less SAR information. In this case, the nearest-neighbor similarity threshold had to be reduced from 0.55 to 0.45 in order to capture more remote similarity relationships. In addition, the neighborhood environment threshold was reduced from 0.4 to 0.3 to increase the number of compounds falling into the neighborhood. Nevertheless, the SPT and the compound potency range are smaller than for the data set in Figure 7a. Despite softening the similarity thresholds, this SPT contains closely related compounds. The SPT is interpretable, and subsets of analogs with increasing potency can be selected. By contrast, in Figure 7c, a fructose-1,6-bisphosphate aldolase inhibitor set of poor SAR information content is analyzed. At first glance, the SPT (that was also calculated with further reduced similarity thresholds) reveals that structurally closely and remotely similar compounds show only little potency variation.

Comparison with Cluster Analysis. Hierarchical clustering is typically used to analyze HTS data. Thus, we also compare the results of clustering with SPT analysis of HTS hit sets. All calculations were carried out using the same fingerprint representation. First, complete-linkage hierarchical clustering of the HTS data sets was carried out using the R software package.[31] In these calculations, the maximum distance for a compound contained in one cluster was set to correspond to the SPT neighborhood similarity threshold. However, under these clustering conditions, a total of 1089 clusters were produced including 539 singletons, yielding an average cluster size of three and a maximal size of 135 compounds. Hence, for direct comparison, the clustering results remained largely inconclusive. Therefore, we next applied Ward's clustering[32] in order to balance the cluster composition and to generate clusters containing active compounds that were comparable in size to SPTs. The resulting dendrogram for the cytochrome P450 3A4 data set containing 24 clusters is shown in Supporting Information, Figure S2. SPTs were compared with clusters containing their root compounds. Generally, we observed that SPT and cluster compositions differed. A representative example is shown in Figure 8. Here the members of the cluster (78 compounds) containing the root compound of the highly ranked SPT shown in Figure 7a (128 compounds) were mapped on the SPT representation (blue nodes, 44 compounds). As can be seen, the cluster covered only a subset of the nearest-neighbor relationships captured by the SPT. Additional compounds contained in this cluster fell outside the SPT neighborhood. Thus, analysis of the clustering and SPT results for this compound focuses on different compound subsets and the results are thus not directly comparable. We note that SPTs are designed to capture nearest-neighbor relationships between individual compounds within a given subset, whereas clustering selects compound subsets without organizing member relationships. Furthermore, clustering captured only a subset of nearest-neighbor relationships contained in an SPT, as illustrated in Figure 8, and only a part of the corresponding SAR information. Moreover, the SPT ranking scheme prioritizes compound subsets on the basis of potency variations among nearest neighbors.

Practical Guidelines. The SPT scoring function is applied to rank SPTs for any data set according to SAR information content. SPT scores are only considered for ranking, i.e., on a relative scale, and their absolute magnitude has no further meaning. After a ranking has been computed, highly ranked SPTs can be quickly scanned to provide an overview of characteristic horizontal and vertical patterns. If data sets contain interpretable SAR information, then this will, in our experience, already become obvious by inspecting only on the order of 10−20 highly ranked SPTs. High-scoring SPTs will often overlap if the selected compound subsets display high SAR information content. However, data sets usually contain different local SARs, and it is therefore advisible to filter SPTs for redundant compounds and also inspect highly ranked nonredundant SPTs in order to explore diverse SAR information. If interpretable horizontal and vertical patterns are limited in highly ranked SPTs, then the probability that a data set contains substantial SAR information is very low. In such cases, as a control, we also suggest inspecting ∼10 nonredundant SPTs. If no patterns become apparent in these SPTs, then it can be concluded that the data contain no detectable SAR information, and the analysis can be terminated. However, we find that the majority of screening data sets we inspect contain SAR information of varying degrees

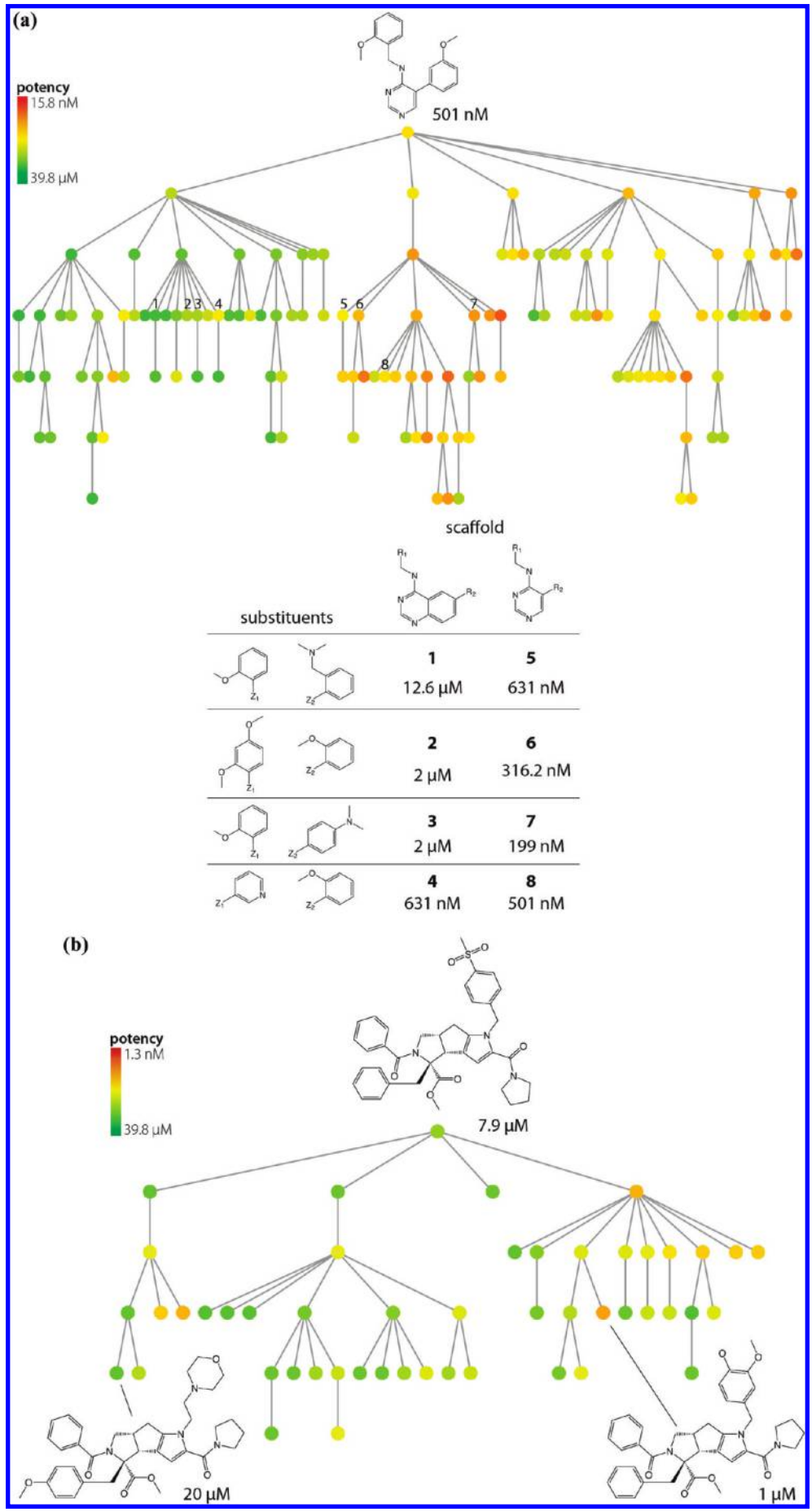
(a)
potency
15.8 nM
39.8 μM
501 nM
1
2
3
4
5
6
7
8
scaffold
substituents
1
12.6 μM
5
631 nM
2
2 μM
6
316.2 nM
3
2 μM
7
199 nM
4
631 nM
8
501 nM
(b)
potency
1.3 nM
39.8 μM
7.9 μM
20 μM
1 μM

Figure 7

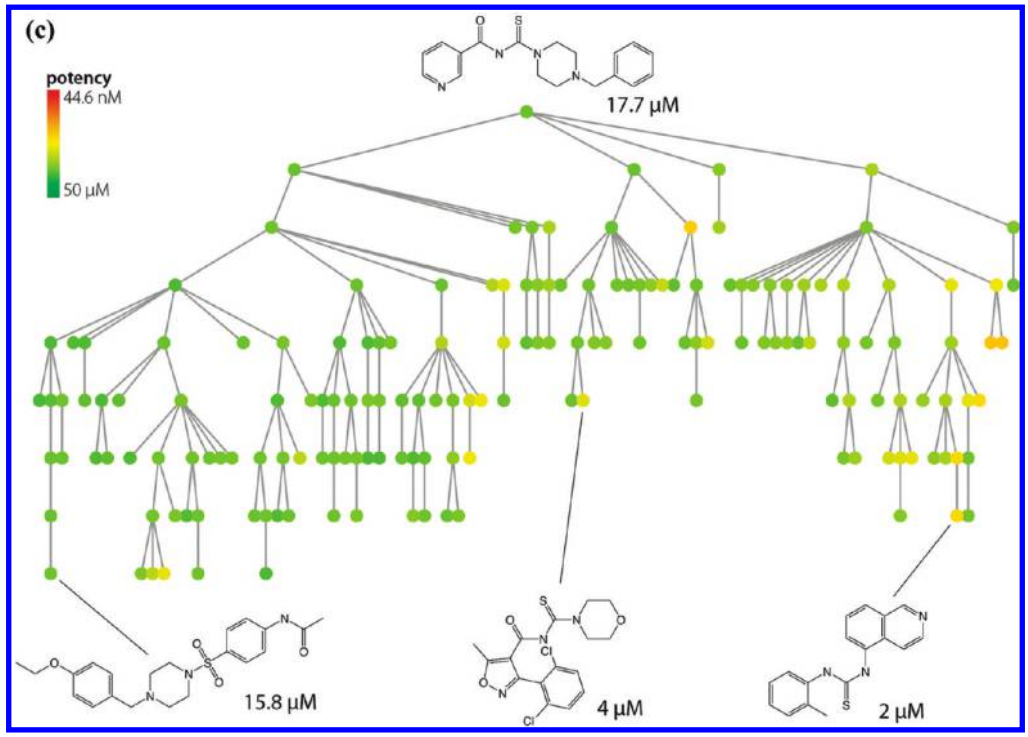

Figure 7. SPT analysis of HTS data sets. In (a), a highly ranked SPT for a set of cytochrome P450 3A4 screening hits is shown. The SPT displays characteristic horizontal and vertical patterns indicating the presence of interpretable SAR information. A subset of two analog series with corresponding substituents but different scaffolds is shown. (b) A highly ranked SPT of thyroid stimulating hormone receptor screening hits is shown that also contains regular horizontal patterns but less SAR information than the SPT in (a). Selected compounds are shown. In (c), the top-ranked SPT of a fructose-1,6-bisphosphate aldolase (from *Giardia lamblia*) screening set is shown that contains only little SAR information. Selected compounds and their potency values are also shown.

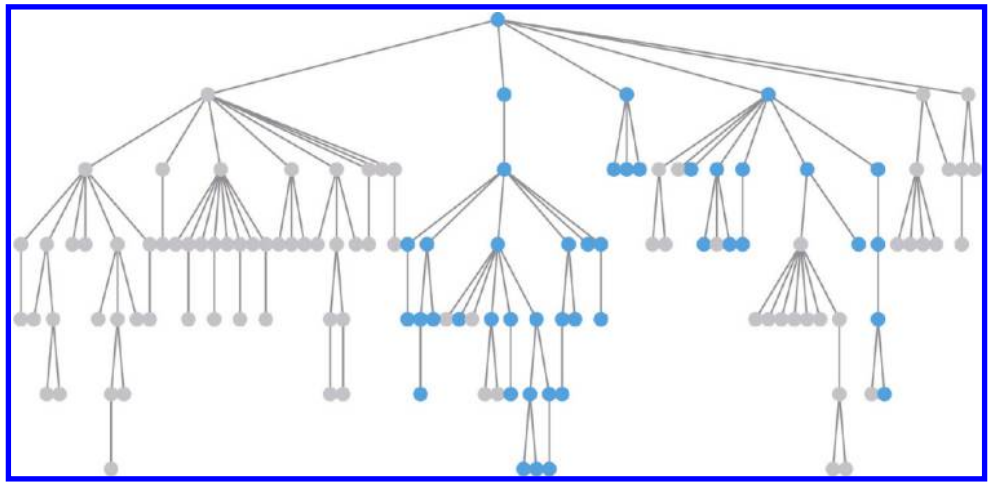

Figure 8. Comparison of SPT and cluster analysis. Shown is the SPT for the cytochrome P450 3A4 HTS set according to Figure 7a without potency coloring. A subset of nodes is colored blue. The corresponding compounds were placed in the same cluster as the root compound when the data set was subjected to hierarchical clustering. The SPT contains 128, and the cluster contains 78 compounds. The overlap between them is 44 compounds. Thus, cluster and SPT analysis yielded only partly overlapping compound subsets, and clustering accounted only for a part of the compound relationships captured in the SPT.

that is easily detectable in highly ranked redundant and nonredundant SPTs.

CONCLUSIONS

The similarity−potency tree (SPT) method is designed to search for available structure−activity relationship (SAR) information in compound data sets of any source and provides an intuitive graphical access to this information. Herein we have introduced the methodology and illustrated characteristic features of the approach using different sets of publicly available enzyme inhibitors. Representative examples have been discussed in order to demonstrate how SAR information can be mined and how SAR rules are deduced. Dividing global data analysis into a series of local tasks is at the core of the approach. SPTs of compound neighborhoods represent a unique data structure that is simple, provide an immediate view of SAR information content, and can be readily interpreted by combining horizontal and vertical graph reading. We find that the availability of interpretable SAR information often depends on the choice of the root compound, even if compound neighborhoods closely overlap. Accordingly,

the ability to characterize overlapping compound neighborhoods and to focus on SPTs that reveal SAR information in most obvious ways is a key feature of the approach. Different SPTs also reveal complementary SAR information, in particular, if they are focused on different compound series. Given the simplicity of the approach and the limited input information that is required, we anticipate that SPTs are widely applicable in large-scale SAR analysis. Importantly, the approach does not build SAR models of predefined structure and does not attempt to replace chemical interpretation or knowledge. Rather, it is a data-driven approach conceptualized to provide an easy access to SAR information and enable interactive interpretation. Upon online publication, the SPT program is made freely available and can be obtained via the following URL (under the "downloads" section): http://www.lifescienceinformatics.uni-bonn.de.

ACKNOWLEDGMENT

The authors would like to thank Lisa Peltason, Peter Haebel, and Nils Weskamp for helpful discussions. M.W. is supported by Boehringer Ingelheim.

Supporting Information Available: Figures S1−S2 show the results of SPT control calculations and a comparison of SPT analysis with hierarchical clustering. This information is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Kubinyi, H. Similarity and Dissimilarity. A medicinal chemist's view. *Perspect. Drug Discovery Des.* **1998**, *9−11*, 225–252.
(2) Gribbon, P.; Lyons, R.; Laflin, P.; Bradley, J.; Chambers, C.; Williams, B. S.; Kighley, W.; Sewing, A. Evaluating Real-Life High-Throughput Screening Data. *J. Biomol. Screening* **2005**, *10*, 99–107.
(3) Kubinyi, H. QSAR and 3D QSAR in Drug Design. Part 1: Methodology. *Drug Discovery Today* **1997**, *2*, 457–467.
(4) 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.
(5) Malo, N.; Hanley, J. A.; Cerquozzi, S.; Pelletier, J.; Nadon, R. Statistical Practice in High-Throughput Data Analysis. *Nat. Biotechnol.* **2006**, *24*, 167–175.
(6) Ahlberg, C. Visual Exploration of HTS Databases: Bridging the Gap between Chemistry and Biology. *Drug Discovery Today* **1999**, *4*, 270–485.
(7) Roberts, G.; Myatt, G. J.; Johnson, W. P.; Cross, K. P.; Blower, P. E., Jr. Lead Scope: Software for Exploring Large Sets of Screening Data. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1302–1314.
(8) Stahl, M.; Mauser, H. Database Clustering with a Combination of Fingerprint and Maximum Common Substructure Methods. *J. Chem. Inf. Model.* **2005**, *45*, 542–548.
(9) Böcker, A. Toward an Improved Clustering of Large Data Sets Using Maximum Common Substructures and Topological Fingerprints. *J. Chem. Inf. Model.* **2008**, *48*, 2097–2107.
(10) Agrafiotis, D.; Shemanarev, M.; Connolly, P.; Farnum, M.; Lobanov, V. SAR Maps: A New SAR Visualization Technique for Medicinal Chemists. *J. Med. Chem.* **2007**, *50*, 5926–5937.
(11) Kolpak, J.; Connolly, P. J.; Lobanov, V. S.; Agrafiotis, D. K. Enhanced SAR Maps: Expanding the Data Rendering Capabilities of a Popular Medicinal Chemistry Tool. *J. Chem. Inf. Model.* **2009**, *49*, 2221–2230.
(12) Wilkens, S. J.; Janes, J.; Su, A. I. HierS: Hierarchical Scaffold Clustering Using Topological Chemical Graphs. *J. Med. Chem.* **2005**, *48*, 3182–93.
(13) Cho, S. J.; Sun, Y. Visual exploration of structure-activity relationship using maximum common framework. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 571–8.
(14) Shanmugasundaram, V.; Maggiora, G. M. Characterizing Property and Activity Landscapes Using an Information-Theoretic Approach. Proceedings of the 222nd American Chemical Society National Meeting, Division of Chemical Information, Chicago, IL, August 26−30, 2001; American Chemical Society: Washington, D.C., 2001; abstract no. *77*.
(15) Peltason, L.; Bajorath, J. SAR Index: Quantifying the Nature of Structure-Activity Relationships. *J. Med. Chem.* **2007**, *50*, 5571–5578.
(16) Guha, R.; Van Drie, J. H. Structure-Activity Landscape Index: Identifying and Quantifying Activity Cliffs. *J. Chem. Inf. Model* **2008**, *48*, 646–658.
(17) Bajorath, J.; Peltason, L.; Wawer, M.; Guha, R.; Lajiness, M. S.; Van Die, J. H. Navigating Structure-Activity Landscapes. *Drug Discovery Today* **2009**, *14*, 698–705.
(18) 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.
(19) Wawer, M.; Sun, S., Bajorath, J. Computational Characterization of SAR Microenvironments in High-Throughput Screening Data. *Intl. J. High Throughput Screen.* **2010**, *1*, 15−27.
(20) Maggiora, G. M. On Outliers and Activity Cliffs - Why QSAR often Disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535–1535.
(21) Wawer, M.; Peltason, L.; Bajorath, J. Elucidation of Structure-Activity Relationship Pathways in Biological Screening Data. *J. Med. Chem.* **2009**, *52*, 1075–1080.
(22) Wawer, M.; Bajorath, J. Systematic Extraction of Structure-Activity Relationship Information from Biological Screening Data. *ChemMedChem* **2009**, *4*, 1431–1438.
(23) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: a Web-Accessible Database of Experimentally Determined Protein-Ligand Binding Affinities. *Nucleic Acids Res.* **2007**, *35*, D198–D201.
(24) PubChem; National Center for Biotechnology Information: Bethesda, MD; http://pubchem.ncbi.nlm.nih.gov. Accessed January 05, 2010.
(25) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
(26) Rogers, D.; Brown, R. D.; Hahn, M. Using Extended-Connectivity Fingerprints with Laplacian-modified Bayesian analysis in High-Throughput Screening Follow-up. *J. Biomol. Screening* **2005**, *10*, 682–686.
(27) *Scitegic Pipeline Pilot, Student ed.*; version 6.1; Accelrys, Inc.: San Diego, CA, 2007.
(28) O'Madadhain, J.; Fisher, D.; White, S. *Java Universal Network/Graph Framework*; Berkeley Software Distribution: University of California, Berkeley; http://jung.sourceforge.net/. Accessed January 11, 2010.
(29) Fry, B. Reas, C. Processing, version 1.0, 2009; http://processing.org. Accessed January 11, 2010.
(30) Rusinko, A., III; Farmen, M. W.; Lambert, C. G.; Brown, P. L.; Young, S. S. Analysis of a large structure/biological activity data set using recursive partitioning. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1017–26.
(31) R Development Core Team; *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2009.
(32) Ward, J. H. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.* **1963**, *58*, 236–244.

CI100197B