ACS Medicinal Chemistry Letters

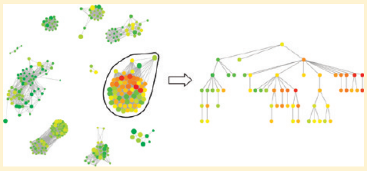
Extracting SAR Information from a Large Collection of Anti-Malarial Screening Hits by NSG-SPT Analysis

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

S Supporting Information

ABSTRACT: We combine two graphical SAR analysis methods, Network-like Similarity Graphs (NSGs) and Similarity-Potency Trees (SPTs), to search for SAR information in a large and heterogeneous compound data set containing more than 13,000 antimalarial screening hits that was recently released by GlaxoSmithKline (GSK). The NSG-SPT approach first identifies subsets of compounds inducing local SAR discontinuity in data sets and then extracts available SAR information from these subsets in a graphically intuitive manner. Applying the NSG-SPT analysis scheme, we have identified in the GSK collection compound subsets of high local SAR information content including both known and previously unknown antimalarial chemotypes, which yielded interpretable SAR patterns. This information should be helpful to prioritize and select antimalarial candidate compounds for further chemical exploration. Furthermore, the NSG-SPT tools are publicly available, and our study also shows how to practically apply these SAR analysis methods to study large compound data sets.



KEYWORDS: Anti-malaria screening hits, data mining, structure−activity relationship (SAR) information, graphical SAR analysis, network-like similarity graphs, similarity-potency trees

We have introduced different numerical and graphical analysis methods to systematically search for SAR information in large compound data sets and extract available information.[1−3] These SAR data mining methods conceptually depart from commonly used statistical, graphical, and molecular classification approaches.[4−10] Among other SAR features, our methods focus on the identification of local SAR environments that are rich in interpretable structure−activity information.[1−3] In order to study relationships between global and local SAR features, Network-like Similarity Graphs (NSGs) have been introduced that represent an annotated similarity-based molecular network structure integrating different levels of SAR-relevant chemical information.[11] Furthermore, the Similarity-Potency Tree (SPT) is another recently introduced data structure designed to explore local SAR environments that only utilizes compound potency values and nearest neighbor similarity relationships.[12] These two data analysis methods are combined herein to identify and characterize SAR environments in a large screening set.

Figure 1 summarizes key features of NSGs and SPTs. In NSGs, compounds are represented as nodes and connected by edges if their calculated pairwise 2D similarity exceeds a predefined threshold value. It should be noted that the assessment of compound similarity is generally influenced by chosen molecular representations (descriptors) and similarity metrics (see Computational Procedures). Nodes are color-coded using a continuous color spectrum reflecting the potency range in a compound set, i.e. from green (lowest) to red (highest potency). Furthermore, nodes are scaled in size according to the contribution of individual compounds to local SAR discontinuity, as assessed by a numerical SAR analysis function.[11] This numerical function systematically relates compound similarity values and potency differences to each other and quantifies SAR contributions on a per-compound basis. Higher values (and thus larger nodes) indicate compounds whose potency significantly deviates from that of their structural neighbors. Thus, pairs of large red and green nodes connected by an edge represent structurally similar compounds with large potency differences that form “activity cliffs”,[1,3] i.e. regions of highest local SAR discontinuity. In addition to accounting for pairwise compound similarity relationships, the data set is also clustered to provide an additional level of similarity information. For visual analysis, a graphical layout algorithm is applied that places multiple densely connected compounds in close vicinity in the graphical representation and separates weakly connected regions from each other.[11] In general, compound subsets that predominantly consist of small similarly colored nodes form continuous local SARs, whereas compounds in clusters with larger red and green nodes form discontinuous local SARs.

SPTs represent a treelike data structure that accesses SAR information in a manner that is complementary to NSGs. A characteristic feature of SPTs is that each SPT is centered on an individual root molecule. Here, compounds are also represented as nodes and are color-coded in analogy to NSGs (Figure 1). However, in SPTs, edges account for 2D structural nearest neighbor relationships (i.e., a compound is only connected to its nearest neighbors), and compounds are arranged in layers of

Received: October 4, 2010
Accepted: December 22, 2010
Published: January 05, 2011

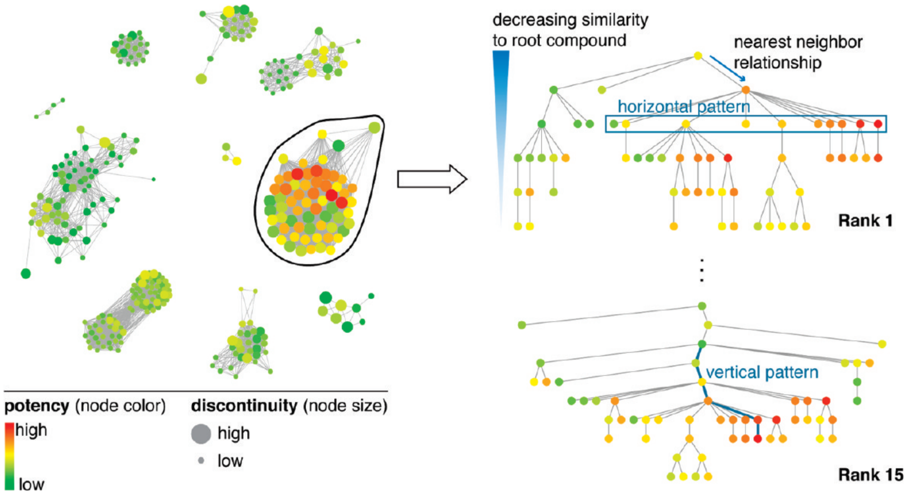

decreasing similarity to root compound
nearest neighbor relationship
horizontal pattern
Rank 1
vertical pattern
Rank 15
potency (node color)
high
low
discontinuity (node size)
high
low

Figure 1. NSG-SPT analysis scheme. On the left, an exemplary NSG is shown calculated for a set of known thrombin inhibitors. This NSG consists of several components that display different local SARs. Nodes represent individual compounds that are connected by edges if they share 2D similarity above a predefined threshold. The color and size of a node reflects the potency and contribution to the local SAR discontinuity of the corresponding compound, respectively, as indicated below the graph. The highlighted region (compound subset) forms the most discontinuous local SAR and was subjected to SPT analysis. For this purpose, each compound is selected once as the root to build a set of overlapping trees. In each SPT, the remaining compounds are connected to the root on the basis of nearest neighbor similarity relationships. Two exemplary SPTs are shown on the right. These SPTs reveal horizontal and vertical SAR patterns that are highlighted. SPTs are ranked based on the occurrence of such patterns.

subsequent nearest neighbor relationships. From the top to the bottom in an SPT, the similarity to the root compound decreases. Systematically generated SPTs represent in part overlapping compound neighborhoods focused on one root compound at a time (hence, for a data set with *n* compounds, *n* trees are obtained). Informative SPTs contain characteristic horizontal and vertical node patterns with systematic potency increases. Horizontal node patterns are formed by different neighbors of the same compound, and vertical patterns, by compounds with subsequent nearest neighbor relationships. SPTs are ranked for further analysis according to their SAR information content using a scoring function that emphasizes well-defined SAR patterns produced by similar compounds with different potency.

NSG-SPT analysis combines compound network and tree analysis in a sequential manner. As illustrated in Figure 1, *NSG-SPT analysis searches for local regions (compound subsets) of significant SAR discontinuity in NSGs and then analyzes the SAR information that is associated with these regions in a compound-centric manner using SPTs.*

The release of the GlaxoSmithKline (GSK) antimalarial screening data[13] has been a pioneering effort, making a large body of compound data generated in the pharmaceutical industry available to the public. It provides a significant opportunity for further discovery efforts focusing on neglected diseases in publicly supported research environments and, in addition, for the development and validation of screening data and SAR analysis methods. The GSK data set contains a total of 13,533 compounds, each of which displayed at least 80% confirmed inhibitory activity in parasite growth assays at 2 μM concentration. Analysis of these hits revealed that they represented 416 different chemotypes,[13] which were defined based on the presence of unique heteroatom scaffolds (and we adhere to this definition herein). Furthermore, target predictions were also carried out for these compounds, suggesting that they might act against as many as 146 different microbial targets.[13] Hence, this hit set is not only large in size but also highly heterogeneous in terms of compound structure and function(s). Thus, extracting SAR information from such data, if available, is a highly relevant but certainly far from routine task. As such, this compound data set represents a prime example for the application of SAR data mining and analysis methods.

Figure 2a shows an NSG of the entire GSK data set. The NSG is interactively navigated. Zooming enables the analysis of graph details, and nodes are graphically associated with corresponding compound structures. Pairwise compound similarity relationships are indicated by gray edges. For clarity, only compounds connected by edges are displayed. This was done not only because of the large data set size but also because compounds without structurally similar neighbors contribute only very little, if any, SAR information. For the purpose of our analysis, one aspect of the standard NSG format[11] was modified, i.e. no clustering was carried out to complement the information provided by the pairwise similarity-based network. This was done because we also compared the results of NSG analysis to conventional cluster analysis of screening data, as further discussed below.

The NSG in Figure 2a illustrates the structural heterogeneity of the GSK data set. There are densely connected central regions in the NSG but also many peripheral nodes with only few or single connections. In the NSG, two regions of prominent local SAR discontinuity become immediately apparent that are highlighted in Figure 2a. These regions are characterized by the presence of many larger red and green/yellow nodes and are shown in detail in Figure 2b. Compounds within these encircled regions in Figure 2a were then selected from the NSG and subjected to SPT analysis. Table S1 of the Supporting Information reports the compound composition of the selected regions and subsequently selected regions, as discussed below. For each region, SPTs were systematically generated, with each compound used once as a tree root, and informative SPTs were analyzed in detail. Representative examples are shown in Figure 2c. The first SPT organizes compounds from NSG region 1 in Figure 2b and contains several compound series that share the same ancestor and cover a broad potency range. Additionally, highly and weakly

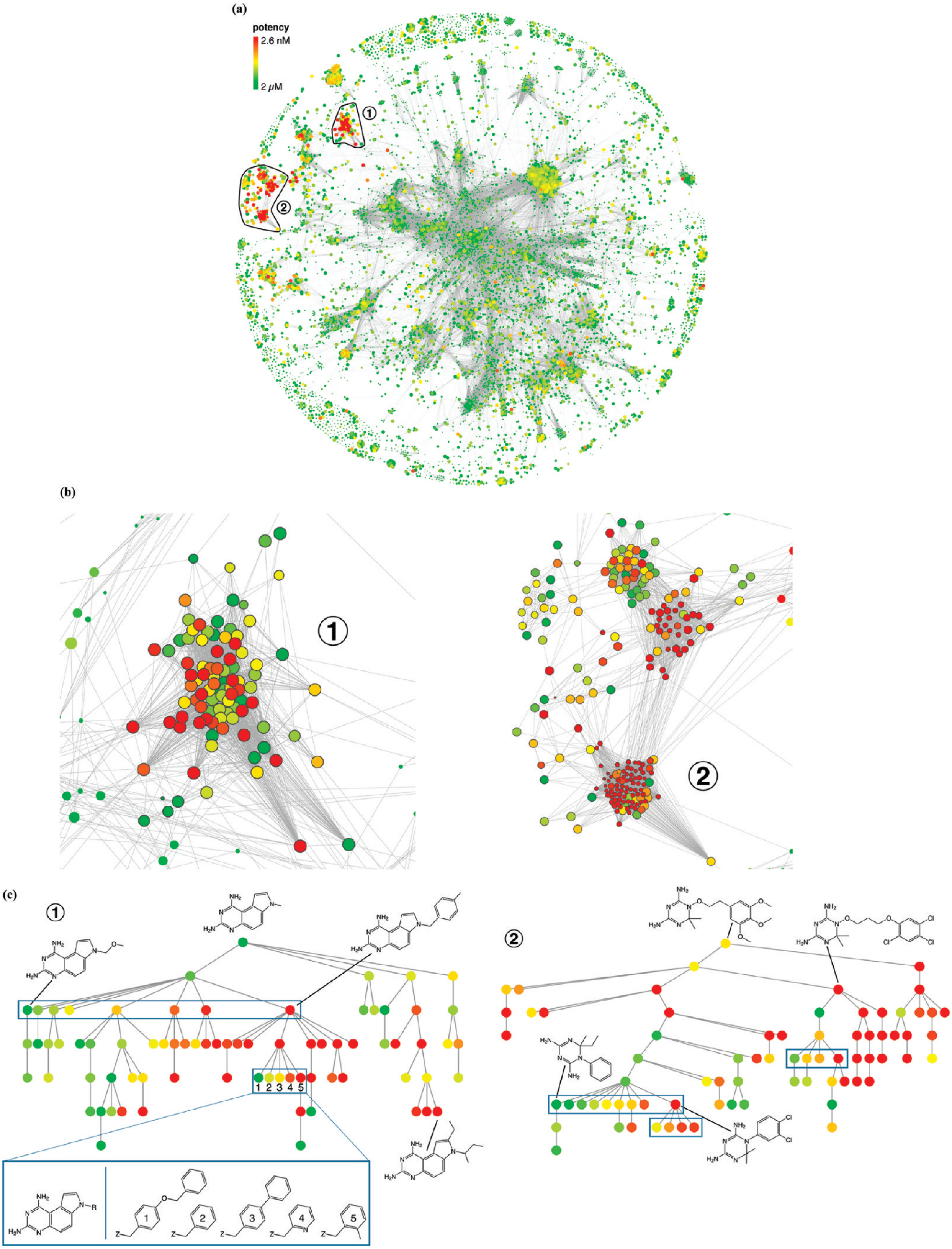


Figure 2. NSG-SPT analysis of the GSK data set. (a) NSG of the complete hit set. Two prominent regions of local SAR discontinuity are highlighted and shown in detail in part b. For these regions, corresponding highly ranked SPTs are provided in part c. Selected compounds are shown, and patterns that reflect significant SAR information are highlighted.

potent compounds form subgroups within the tree and generate an ordered potency distribution. Horizontal patterns of structurally similar compounds (analog series) with gradually increasing potency are evident and highlighted in Figure 2c. Hence, this

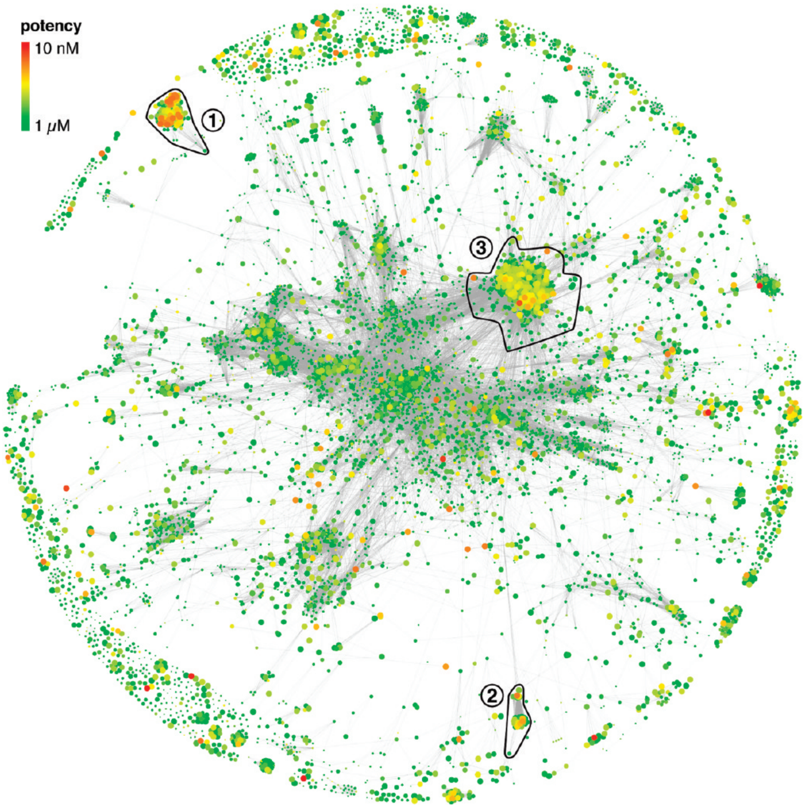


Figure 3. NSG-SPT analysis after removal of known antimalarial chemotypes. NSG of the GSK data set after removal of known anti-malarial chemotypes. The positions of the remaining nodes correspond to the layout in Figure 2a. To account for the removal of highly potent compounds, the potency-based coloring was adjusted to range from 1 μM (green) to 10 nM (red).

SPT and also other overlapping SPTs reveal clear SAR information. Similar patterns can be observed in the second SPT in Figure 2c, which contains compounds from NSG region 2 in Figure 2b. The compounds in both SPTs are derivatives of or structurally related to diaminopyrimidines and diaminotriazines that dominate the corresponding NSG regions and are already known classes of antimalarial compounds. Hence, at least some of the clear SAR patterns we observed in our SPTs were likely the result of optimization efforts focusing on this compound class. This SAR information was most prominently displayed by the GSK data set.

In order to focus our subsequent analysis on previously unknown inhibitory chemotypes, we collected known antimalarial compounds from major public domain repositories of bioactive compounds, BindingDB[14] and ChEMBL,[15] and examined whether these known compounds or similar molecules were present in the GSK collection. In total, we identified 2914 known antimalarial compounds in the data set that corresponded to 1186 different chemotypes and removed these compounds from the GSK set. We then recalculated the NSG for the remaining data set, as shown in Figure 3. For comparison with Figure 2a, we retained the NSG layout computed for the original data set such that removal of known active compounds created “holes” in the NSG in Figure 3. However, due to the removal of highly potent known compounds, the potency distribution in the data set was modified, and we thus adjusted the color code to span a potency range from 1 μM to 10 nM. This adjustment further emphasized other local NSG regions of notable SAR discontinuity. Three selected regions (1−3) are highlighted in Figure 3 and are shown in detail in Figure S1 of the Supporting Information. Here, local SAR discontinuity was more characteristic of screening hits, because known highly potent compounds were absent that dominated the NSG regions displayed in Figure 2b. However, the first compound subset in Figure S1 of the Supporting Information also displays a high degree of SAR discontinuity, and the corresponding SPT in Figure S2 of the Supporting Information shows a fairly ordered potency distribution with horizontal and vertical patterns that largely separate weakly potent and moderately to highly potent compounds from each other and reveal SAR trends. For example, the right branch of the SPT contains series of similar compounds with varying potency that can be readily selected for further analysis. Region 2 in Figure S1 of the Supporting Information contains comparably few compounds, among them a series of close analogues. In the corresponding SPT in Figure S2 of the Supporting Information, several analogues with medium potency are found to have child nodes with further increasing potency. Different from region 2, region 3 in Figure S1 of the Supporting Information is densely populated, including very many weakly potent compounds that do not contribute to local SAR discontinuity (small green nodes). However, there are also numerous larger yellow to orange and red nodes that induce a notable degree of local discontinuity (that is, at first glance, masked by the large number of structurally similar background compounds). The corresponding SPT in Figure S2 of the Supporting Information is also rather dense, and for clarity, only the right branch of the SPT is shown (the complete SPT is provided in Figure S3 of the Supporting Information). The right branch of the SPT reveals a series of compounds that display vertical patterns of potency variations, and these compounds would also be attractive candidates for further analysis.

Interactive analysis of the NSG in Figure 3 reveals a variety of other small regions with compounds containing apparent SAR information that can be analyzed analogously to the representative examples discussed above.

Although compound clustering provides a standard information layer of NSGs, it was deliberately omitted here from NSG generation, as mentioned above. Thus, compound selection only on the basis of a pairwise similarity network annotated with per-compound SAR discontinuity contributions could be directly compared to cluster analysis of the data set. Therefore, the reduced GSK data set (without previously known antimalarial chemotypes) was subjected to k-means clustering and divided into 60 nonoverlapping clusters. The cluster and potency distribution is reported in Figure S4 of the Supporting Information. The composition of all clusters is freely available via the following URL: http://www.lifescienceinformatics.uni-bonn.de (please, see the Downloads section). Clusters with highest median compound potency and clusters containing the most potent remaining compounds were selected and mapped onto the NSG representation, as shown in Figure S5 of the Supporting Information. As can be seen, the selected clusters overlap with regions that were prioritized on the basis of NSG analysis. Hence, the NSG network layout already accounted for compound clustering effects, without the addition of an explicit clustering step. In this case, clusters selected on the basis of highest median potency also contained compound subsets that displayed SAR trends. However, potency-oriented cluster selection cannot replace a systematic account of pairwise similarity and potency relationships for large-scale SAR analysis. For example, in heterogeneous data sets containing chemically different series of highly potent compounds, potency-oriented compound selection is likely to overlook the presence of SAR patterns at the midpotency range that are often of particular interest for further chemical exploration.

In summary, we have screened the large and heterogeneous GSK antimalarial data set for SAR information using graphical analysis tools. We have shown that a screening set containing more than 13,000 hits with likely activity against more than 100 targets can be analyzed in a graphically intuitive manner. A key aspect of NSG-SPT analysis, as presented herein, is that one first focuses on the identification of compound subsets that are discontinuous in their SAR behavior and then extracts available SAR information in detail. Although SAR information in the GSK data set is overall sparsely distributed, as one might expect given its high-throughput screening origin and heterogeneity, different local environments with notable SAR information content have been identified. The most significant SAR information contained in this data set was associated with previously known antimalarial chemotypes. However, in addition, other compound subsets were also found in the GSK collection that displayed interpretable SAR patterns and should merit further evaluation. Thus, the analysis should help to prioritize compound subsets from this large pool of antimalarial screening hits for follow-up studies.

EXPERIMENTAL PROCEDURES

SPTs and NSGs have been generated and represented as described previously.[11,12] As a fingerprint for compound comparison, ECFP4[16] was used. As a criterion for edges between nodes in NSGs, connected compounds needed to exceed a Tanimoto similarity threshold value of 0.4. In SPTs, compounds were only considered nearest neighbors above a similarity threshold value of 0.55. For the third SPT in Figure S2 of the Supporting Information which represents a very dense cluster of similar compounds, the nearest neighbor threshold for SPT generation was raised to 0.7. A previously reported scoring function[12] was used to rank SPTs computed for a compound cluster. This scoring function prioritizes SPTs that contain multiple horizontal and vertical patterns formed by analog series or pairwise similar compounds with gradually increasing potency. Standard *k*-means clustering[17] of the data set was carried out using WEKA.[18] The number of clusters was set to 60, which resulted in clusters with balanced inner-cluster distance distributions. NSG tools are publicly available as a part of the SARANEA software[19] and the SPT program is also available without restrictions via the following URL: http://www.lifescienceinformatics.uni-bonn.de/ (see the Downloads section).

ASSOCIATED CONTENT

Ⓢ **Supporting Information.** Figure S1 shows details of regions selected in Figure 3, and Figure S2 the corresponding SPTs. Figure S3 shows the complete structure of the third SPT in Figure S2. Figure S4 reports the *k*-means cluster and potency distribution for the GSK data set. Figure S5 shows selected clusters mapped on the NSG from Figure 3. Table S1 reports the composition of all clusters of compounds discussed in the text. This material is available free of charge via the internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

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

ACKNOWLEDGMENT

M.W. is supported by Boehringer Ingelheim Pharma, Biberach, Germany.

REFERENCES

(1) 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.

(2) Wawer, M.; Lounkine, E.; Wassermann, A. M.; Bajorath, J. Data Structures and Computational Tools for the Extraction of SAR Information from Large Compound Sets. *Drug Discovery Today* **2010**, *15*, 630–639.

(3) Wassermann, A. M.; Wawer, M.; Bajorath J. Activity Landscape Representations for Structure-Activity Relationship Analysis. *J. Med. Chem.* **2010**, *53*, 8209−8223.

(4) Malo, N.; Hanley, J. A.; Cerquozzi, S.; Pelletier, J.; Nadon, R. Statistical Practice in High-Throughput Data Analysis. *Nat. Biotechnol.* **2006**, *24*, 167–175.

(5) Ahlberg, C. Visual Exploration of HTS Databases: Bridging the Gap between Chemistry and Biology. *Drug Discovery Today* **1999**, *4*, 270–485.

(6) 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.

(7) 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. Screen* **2005**, *10*, 99–107.

(8) Kibbey, C.; Calvet, A. Molecular Property eXplorer: A Novel Approach to Visualizing SAR Using Tree-maps and Heatmaps. *J. Chem. Inf. Model.* **2005**, *45*, 523–532.

(9) Harper, G.; Pickett, S. D. Methods for mining HTS data. *Drug Discovery Today* **2006**, *11*, 694–699.

(10) 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.

(11) 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.

(12) Wawer, M.; Bajorath, J. Similarity-Potency Trees: A Method to Search for SAR Information in Compound Data Sets and Derive SAR Rules. *J. Chem. Inf. Model.* **2010**, *50*, 1395–1409.

(13) Gamo, F.-J.; Sanz, L. M.; Vidal, J.; de Cozar, C.; Alvarez, E.; Lavandera, J.-L.; Vanserwall, D. E.; Green, D. V. S.; Kumar, V.; Hasan, S.; Brown, J. R.; Peishoff, C. S.; Cardon, L. R.; Garcia-Bustos, J. F. Thousands of Chemical Starting Points for Antimalarial Lead Identification. *Nature* **2010**, *465*, 305–312.

(14) 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.

(15) ChEMBL. http://www.ebi.ac.uk/chembl (accessed July 1, 2010).

(16) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

(17) Witten, I. H.; Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed.; Morgan Kaufmann: San Francisco, 2005; pp 136—139.

(18) Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. The WEKA Data Mining Software: An Update. *SIGKDD Explorations* **2009**, *11*, 10–18.

(19) Lounkine, E.; Wawer, M.; Wassermann, A. M.; Bajorath, J. SARANEA: A Freely Available Program to Mine Structure-Activity and Structure-Selectivity Relationship Information in Compound Data Sets. *J. Chem. Inf. Model.* **2010**, *50*, 68–78.