SARANEA: A Freely Available Program To Mine Structure−Activity and Structure−Selectivity Relationship Information in Compound Data Sets

Eugen Lounkine, Mathias Wawer, Anne Mai Wassermann, and Jürgen Bajorath*

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

Received October 26, 2009

We introduce SARANEA, an open-source Java application for interactive exploration of structure−activity relationship (SAR) and structure−selectivity relationship (SSR) information in compound sets of any source. SARANEA integrates various SAR and SSR analysis functions and utilizes a network-like similarity graph data structure for visualization. The program enables the systematic detection of activity and selectivity cliffs and corresponding key compounds across multiple targets. Advanced SAR analysis functions implemented in SARANEA include, among others, layered chemical neighborhood graphs, cliff indices, selectivity trees, editing functions for molecular networks and pathways, bioactivity summaries of key compounds, and markers for bioactive compounds having potential side effects. We report the application of SARANEA to identify SAR and SSR determinants in different sets of serine protease inhibitors. It is found that key compounds can influence SARs and SSRs in rather different ways. Such compounds and their SAR/SSR characteristics can be systematically identified and explored using SARANEA. The program and source code are made freely available under the GNU General Public License.

INTRODUCTION

In medicinal chemistry, structure−activity relationship (SAR) analysis plays a fundamental role in understanding structural determinants of biological activity and in optimizing compounds to become leads for preclinical evaluation.[1,2] Collecting data from hit-to-lead and lead optimization projects makes it possible to study SARs in a comparative manner across different compound classes.[2] The comparative study of SARs is a relatively new field because, traditionally, SARs have mostly been explored for individual compound series on a case-by-case basis.[2,3] Methodologically, comparative SAR analysis requires considering much larger amounts of data than the analysis of individual SARs. Furthermore, one also attempts to extract available SAR information from very large amounts of biological screening data to aid in the selection of hits that are most promising for further chemical exploration.[2] Because individual compounds are often active against multiple related targets,[1,4] comparative SAR analysis represents a multidimensional task that also includes the assessment of compound selectivity arising from differences in potency against multiple targets, leading to the concept of structure−selectivity relationships (SSRs).[4,5]

To facilitate large-scale SAR (and SSR) analysis, we have developed a number of computational methods and tools including the global and local SAR index (SARI),[6,7] network-like similarity graphs (NSGs),[7,8] SAR pathways,[8,9] and SAR trees.[9] These methodologies are designed to elucidate and quantify global and local SAR features contained in any compound data set including raw screening data and identify compounds that determine SAR and/or SSR characteristics.

Given the enormous amount of bioactivity data that have become available for many targets and the increasing complexity of SAR research, there is a growing need in the medicinal chemistry community for intuitive and data-oriented SAR analysis tools.[3] To these ends, we introduce SARANEA, a freely available open-source Java program that allows the interactive and simultaneous exploration of multiple SARs and SSRs. SARANEA integrates our previously reported methodologies[7−9] with new SAR/SSR analysis functions developed by us and presented herein. We have put emphasis on making the tools intuitive and easy to use. They are designed for medicinal and computational chemists but do not require a high level of computational expertise.

Molecular network representations are a central feature of the program, and hence, we have chosen the name SARANEA, which combines "SAR" and "ARANEAE", the scientific designation of the order of spiders, because networks might occasionally remind us of spider webs.

In this paper, we describe in detail the analysis functions implemented in SARANEA and present an exemplary application comparing SARs and SSRs for different sets of serine protease inhibitors, leading to the identification of structural determinants of compound potency and selectivity.

METHODS AND PROGRAM FEATURES

SARANEA Functions. SARANEA integrates different approaches to analyze and compare SARs and SSRs and provides an intuitive graphical user interface. In addition to previously reported global and local SARI scoring, NSGs, SAR pathways, and SAR trees, new analysis functions were developed and implemented in SARANEA. These functionalities include layered chemical neighborhood graphs, cliff indices, and marker compound NSGs. In addition, molecular

* Corresponding author phone: +49-228-2699-306; fax: +49-228-2699-341; e-mail: bajorath@bit.uni-bonn.de.

10.1021/ci900416a
Published on Web 01/07/2010

Table 1. SAR and SSR Analysis Functions in SARANEA[a]

functionality	purpose/description	refs
SARI scores	quantification of global and local SAR and SSR features	5−7;
network-like similarity graphs	interactive visualization of data sets combining compound similarity and potency or selectivity information; selection of compounds from the graph, table view of compound properties, and molecule structure depiction in tooltips	6−7; interactive graphs, editing, structure depiction, and annotation: new functions
synchronized selection	finding key compounds in different data sets and comparing potency and selectivity NSGs	new function
cliff index	prioritization of activity or selectivity cliffs	new function
SAR pathways	identifying sequences of similar compounds that form potency or selectivity gradients	8−9; interactive pathway editing: new function
SAR trees	organization of SAR/SSR pathways that share a compound of interest	8, 9
chemical neighborhood graph	detailed view of the chemical neighborhood of a compound with potency or selectivity information; reveals activity/selectivity cliffs	new function
marker compounds	flagging active compounds using different markers having undesired properties	new function
molecule information	summary of potency and selectivity of a compound across multiple targets	new function

[a] Key functionalities for SAR and SSR analysis implemented in SARANEA are summarized.

networks can be interactively edited, and the compound corresponding to each node in a network can be displayed and its bioactivity summary shown. Table 1 provides an overview of SAR/SSR analysis functions available in SARANEA. The theory underlying these analysis functions and calculation details are provided in the Supporting Information.

Generating Compound Selectivity Sets. Selectivity of a compound for one target over another is defined as the difference in compound potency on a logarithmic scale, e.g., the difference in the pK_i values. This target pair-based calculation yields negative values if the compound is selective for the first target over the second one and positive values for the reverse case. SARANEA automatically assembles selectivity sets containing at least five molecules on the basis of available compound potency information. The target list, which is shown in the main program window, is automatically updated when a target is selected, displaying only those targets that share at least five compounds with the selected one. Selectivity is reported as a normalized ratio of the form "1:*x*" or "*x*:1", dependent on whether the log-scale selectivity value is positive or negative, respectively. For example, a selectivity of "10:1" means that the compound is 10 times more potent for the first target and corresponds to a log-scale selectivity value of −1.

SAR Index. The global SARI is a numerical function designed to assign an SAR category to given sets of active compounds.[6] Possible values range from 0 to 1, and low, intermediate, and high values reflect three general SAR types, discontinuous, heterogeneous, and continuous SARs, respectively.[6] Score calculations are based on compound similarity and potency values. To compare different data sets, SARI scores are normalized with respect to a reference panel of compound data sets representing different SAR categories. A modified version of SARI utilizing potency ratios instead of potency values has been introduced for the classification of SSRs.[5]

Compound Discontinuity Score. To estimate the contribution of individual compounds to SAR discontinuity within a compound set, a compound-based SARI scoring scheme has been introduced, with compound discontinuity scores ranging from 0 to 1.[7] High values indicate high SAR discontinuity, which means that the compound is involved in the formation of "activity cliffs" [2,3] or "selectivity cliffs".[5] Activity cliffs in a data set are formed by structurally similar compounds having dramatic differences in potency. Accordingly, selectivity cliffs are formed by similar compounds with substantial differences in target selectivity. Compound SAR and SSR discontinuity scores are calculated in the same manner using compound potency or target pair selectivity information, respectively. SARANEA normalizes per-compound discontinuity scores on the basis of a panel of reference compound sets extracted from BindingDB,[11] which makes the scores comparable across multiple targets and permits the calculation of normalization parameters for various molecular representations. Details of molecular similarity assessment, the molecular input format, and the normalization procedures are reported in the Supporting Information.

Network-like Similarity Graphs and Associated Features. A graph-based data structure, called network-like similarity graphs, has been introduced for the analysis and visualization of similarity and potency distributions within a compound or screening data set.[7] Each molecule is represented by a node, and nodes are connected by an edge if the structural similarity of the two compounds exceeds a predefined threshold value. The potency range within a compound set is mapped to a continuous node color spectrum from green (lowest potency) to red (highest potency). As an additional layer of information, the compound discontinuity score is reflected by the size of the nodes; i.e., larger nodes correspond to higher discontinuity score values. For visualization purposes, a layout algorithm is applied that clusters densely connected groups of nodes and separates these clusters for clarity. NSG representations can be utilized to analyze both SARs and SSRs.[5,7] In the SSR NSGs, the color code represents compound selectivity for the first (green) or the second (red) target.

SARANEA provides an interactive graph view that permits editing of the graphs including panning, zooming, and selection of individual nodes. The selection can be synchronized across multiple NSG representations, hence making it possible to compare local SAR or SSR environments of selected compounds for different targets. Furthermore, tooltips appear when the mouse is moved over a node showing the corresponding molecular structure and the name and potency or selectivity values associated with the compound. Individual compounds can be selected using their identifier or by drawing a selection box around their nodes using the mouse. Then these compounds are shown in a table that

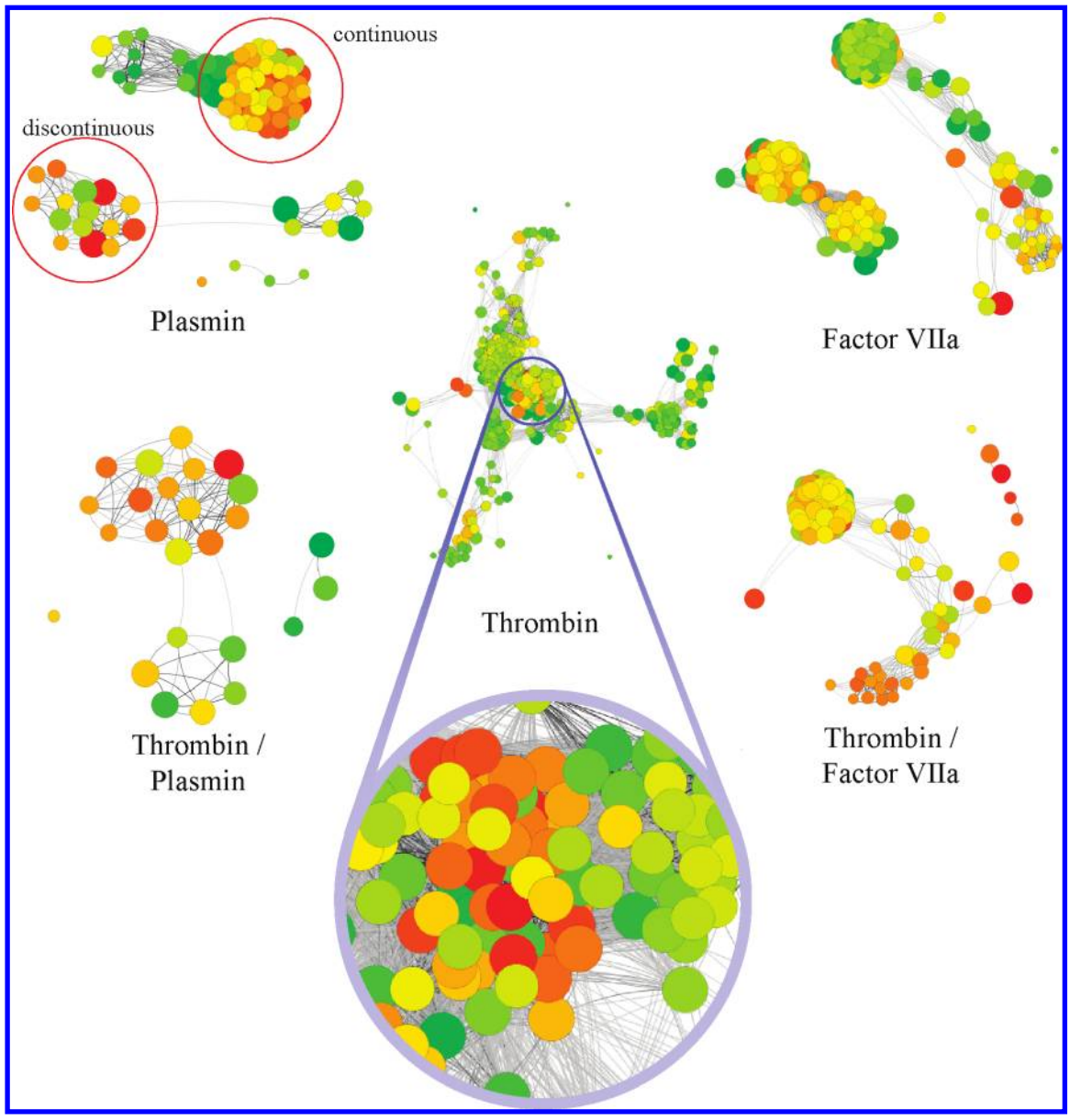
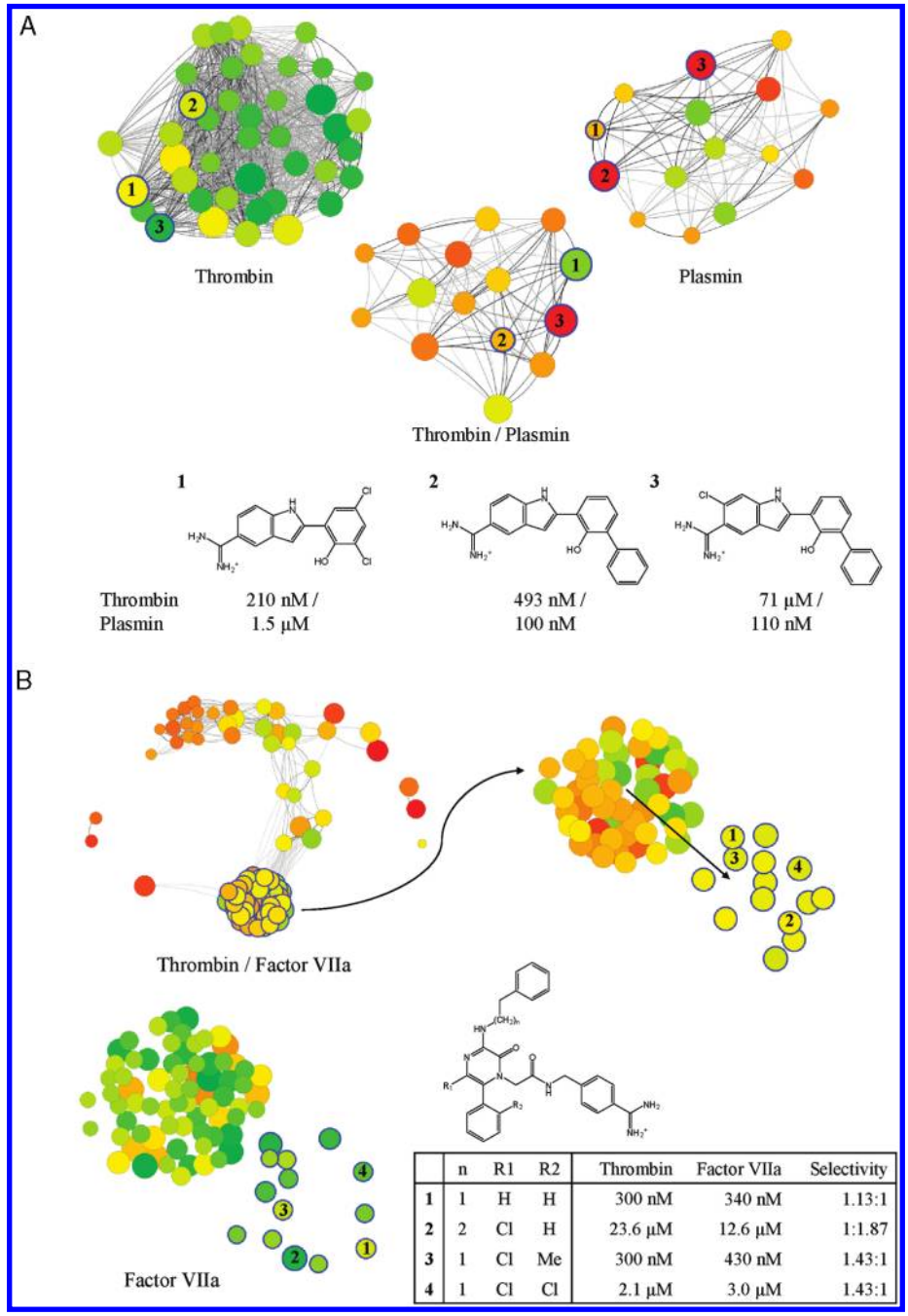


Figure 1. Network-like similarity graphs. Three potency and two selectivity NSGs generated with SARANEA are depicted. Nodes are color-coded on the basis of potency or selectivity. The size of the nodes reflects the compound discontinuity score, and areas containing large red and green nodes represent activity or selectivity cliffs. The selectivity NSGs only contain compounds shared by the two targets that are compared and are therefore smaller than the potency NSGs. In the plasmin NSG (top left), two subsets of compounds are circled that represent continuous and discontinuous local SARs, respectively. The graphs were created using the graph image export function of SARANEA.

reports NSG-specific and user-defined molecular properties. These properties include structural similarity, compound discontinuity scores, and a "cliff index" to prioritize molecule pairs forming activity or selectivity cliffs. The calculation of the cliff index is described in the Supporting Information. The higher the cliff index value for a compound pair, the larger the magnitude of the activity or selectivity cliff they form. For every node, a context menu provides access to additional functions including the selection of structurally similar neighbors or the display of a molecule information window. This window shows available potency and selectivity information of a compound for all targets and target pairs. It can be accessed from the node context menu or by selecting the name of a compound from a drop-down list in the main program window.

Figure 1 shows five exemplary NSGs. The figure was generated utilizing an image export function that is also implemented in SARANEA. Parts A and B of Figure 2 provide examples of compound selection across different NSGs. They are discussed in detail in the Results.

Marker Compounds. For the calculation of NSGs, marker compounds can be added in SARANEA that help to identify active compounds having structural features that might render them problematic. Therefore, molecules that are known to have an undesired property (e.g., carcinogenicity) are added to graph calculations. These markers are connected to active compounds if they exhibit pairwise similarity above the predefined threshold. By selecting a different color code for each marker set, they can be easily distinguished in the NSG representations. For the current implementation of SARANEA, we have assembled organ-specific sets of carcinogenic marker compounds from The Carcinogenic Potency Project (CPDB).[10] A detailed description of these sets can be found in the Supporting Information. Figure 3



	n	R1	R2	Thrombin	Factor VIIa	Selectivity
1	1	H	H	300 nM	340 nM	1.13:1
2	2	Cl	H	23.6 μM	12.6 μM	1:1.87
3	1	Cl	Me	300 nM	430 nM	1.43:1
4	1	Cl	Cl	2.1 μM	3.0 μM	1.43:1

Figure 2. Activity and selectivity cliffs. (A) Three compounds are shown that have different potency against and selectivity for thrombin and plasmin. Compounds 1 and 2 form an activity cliff in the plasmin but not the thrombin inhibitor set. By contrast, compounds 2 and 3 have nearly the same potency against plasmin. However, they form a selectivity cliff, as can be seen in the selectivity NSG (bottom), which is a consequence of different potency against thrombin. (B) illustrates the identification of structural determinants that influence the potency of pyrazinone derivatives against thrombin and factor VIIa. Pyrazinone analogues form a cluster in the thrombin/factor VIIa selectivity NSG (top left) and are selected for further analysis. From the selected subset, compounds with similar potency against the two targets (i.e., nonselective compounds) are isolated (top right). The selected compounds are then analyzed in the factor VIIa potency NSG (bottom left). Here most of these compounds introduce SAR discontinuity.

provides an example of two different marker sets added to an NSG. Active compounds that are connected to markers are flagged for possible carcinogenic effects.

SAR Pathways and Trees. SAR pathways were developed to detect and represent continuous SARs within a data set and aid in the selection of compounds for chemical exploration.[8] On the basis of the NSG data structure, the method searches for sequences of pairwise similar compounds that exhibit a gradual increase in potency along the sequence. SAR pathways can be generated for chosen start and end points, for example, from a compound of interest leading to an activity cliff.[8,9] Figure 4 shows different pathways that evaluate structurally similar compounds in different SAR and SSR contexts. Such comparisons make it possible to identify structural features in compound series that determine their biological activity across multiple targets.

To integrate and compare the information provided by individual SAR pathways, all pathways or a subset of pathways that lead to or originate from a specified compound can also be organized in a treelike structure termed an SAR tree.[9] SARANEA extends the concept of SAR pathways and trees to the study of SSRs and provides an "SARTree" view

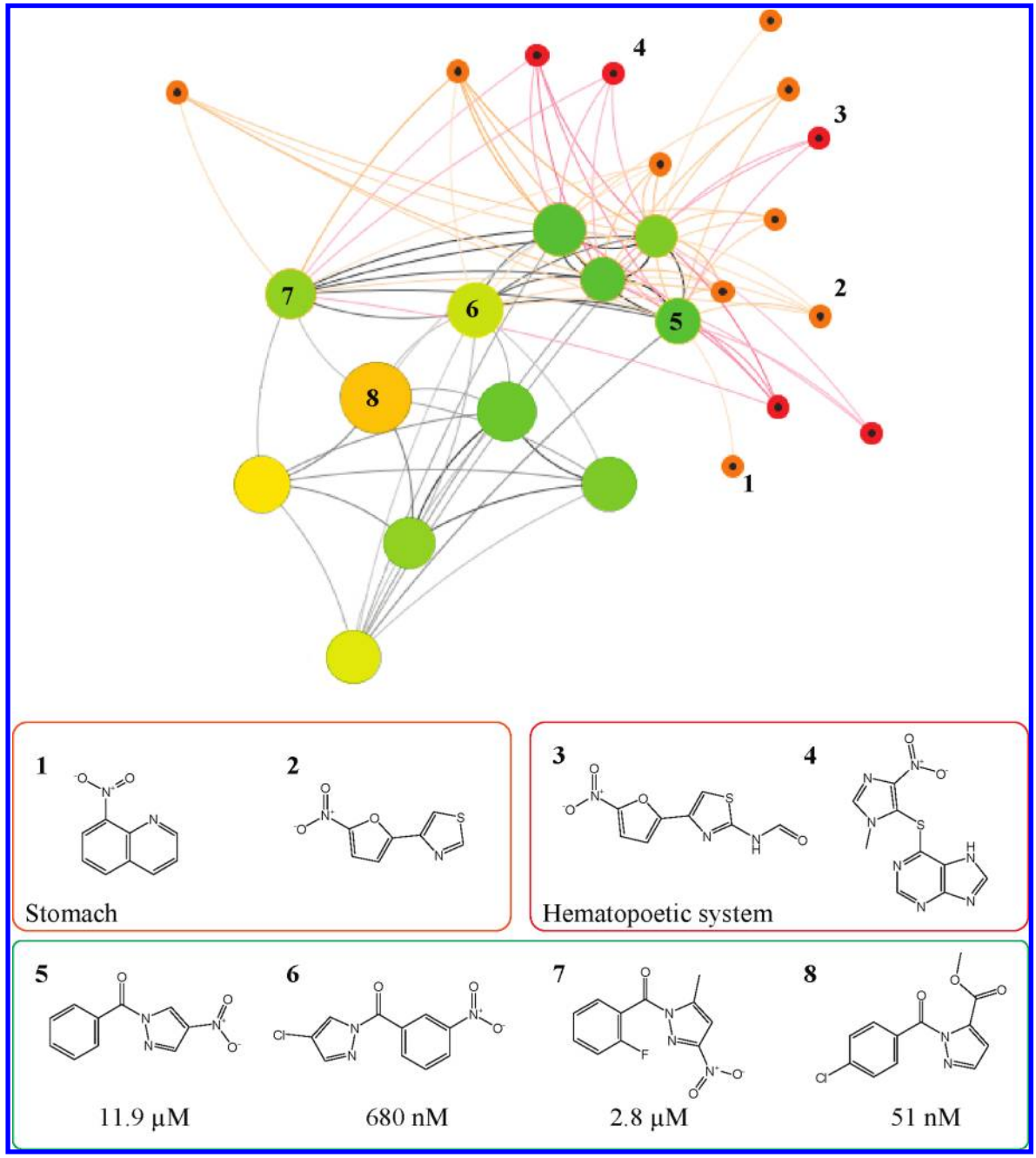


Figure 3. Marker compounds. A subgraph of the thrombin potency NSG is shown following the inclusion of marker compounds that represent organ-specific carcinogenicity sets for the stomach (orange) and hematopoietic system (red). For marker compounds, similarity relationships to thrombin inhibitors are analyzed. The boxes depict exemplary compounds that correspond to numbered nodes in the graph. The orange and red boxes contain molecules from the two marker sets, while the green box shows structurally similar thrombin inhibitors that are immediate or second-degree neighbors of marker compounds.

that shows all pathways containing a specified compound. An exemplary SSR tree is shown in Figure 5. In addition, SARANEA permits interactive modification of pathways based on three complementary functionalities. First, compounds that are neighbors in chemical space can be directly connected, thus eliminating molecules that are positioned between them. These pathway "shortcuts" can be utilized to avoid the inclusion of analogues with only little change in potency or selectivity. Second, pathways can also be further extended by adding compounds between two molecules within the path. For this purpose, for any two compounds A and B, a list of structurally similar molecules can be generated for which potency or selectivity values fall between those of A and B. In NSGs, such molecules are neighbors of both A and B. This feature allows the exploration of additional structural variations that result in an increase in potency or selectivity for compounds that are not selected for a given pathway by the pathway function (see the Supporting Information). Third, layered chemical neighborhood graphs are introduced to organize compounds with decreasing similarity to a central molecule selected from a pathway. Therefore, compounds are projected on concentric circles that represent levels of decreasing similarity relative to the central reference compound. The nodes are color-coded and sorted according to potency or selectivity. An example is provided in Figure 6. The chemical neighborhood graph makes it possible to identify activity and selectivity cliffs, as well as dissimilar compounds having comparable potency

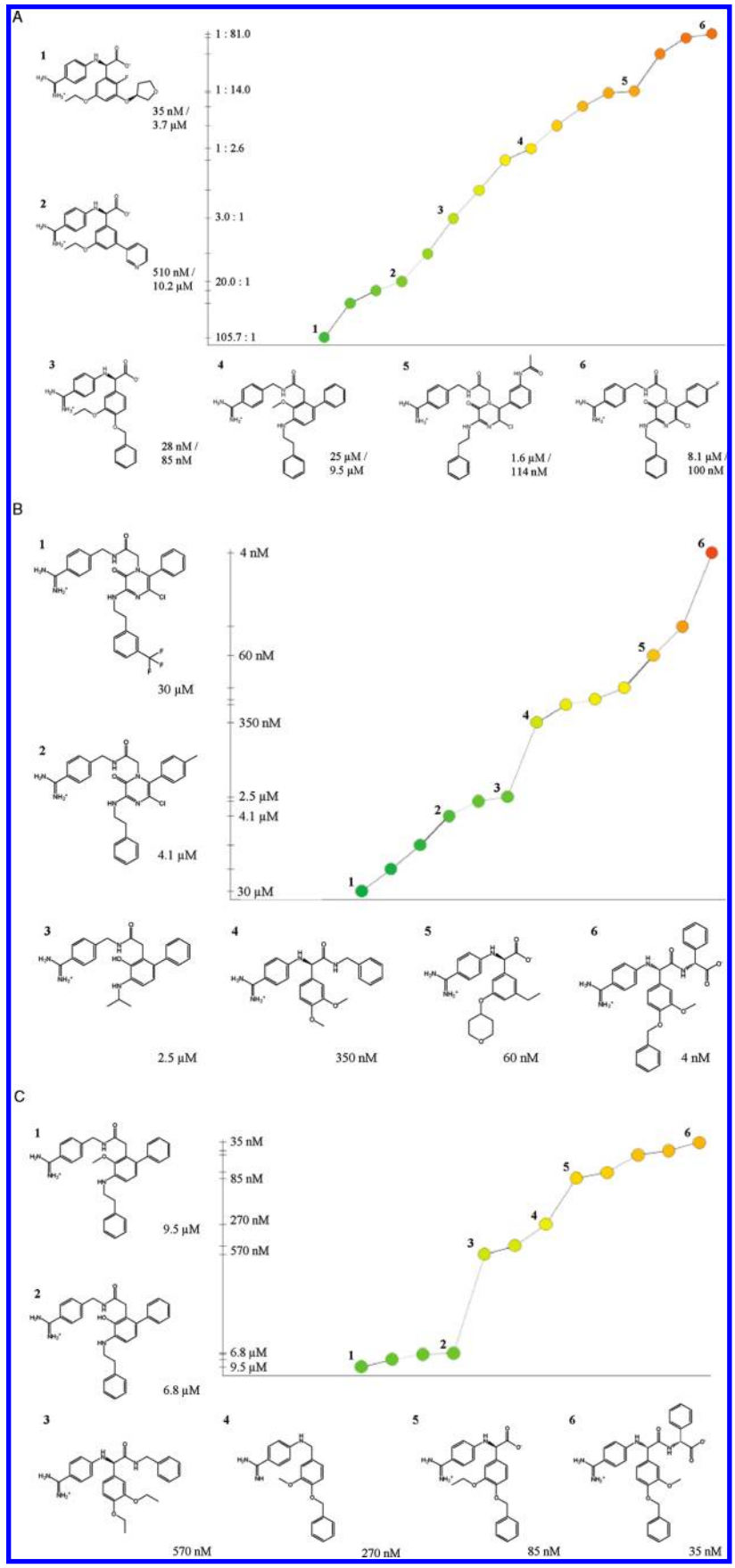


Figure 4. Pathways. (A) shows an SSR pathway for the thrombin/factor VIIa selectivity set. Green nodes represent compounds selective for factor VIIa and red nodes compounds selective for thrombin. Nodes are color-coded and positioned in the graph according to compound selectivity. Molecules corresponding to numbered nodes are shown, and potencies against factor VIIa/thrombin are reported. In (B) and (C), SAR pathways are shown for factor VIIa and thrombin inhibitors, respectively.

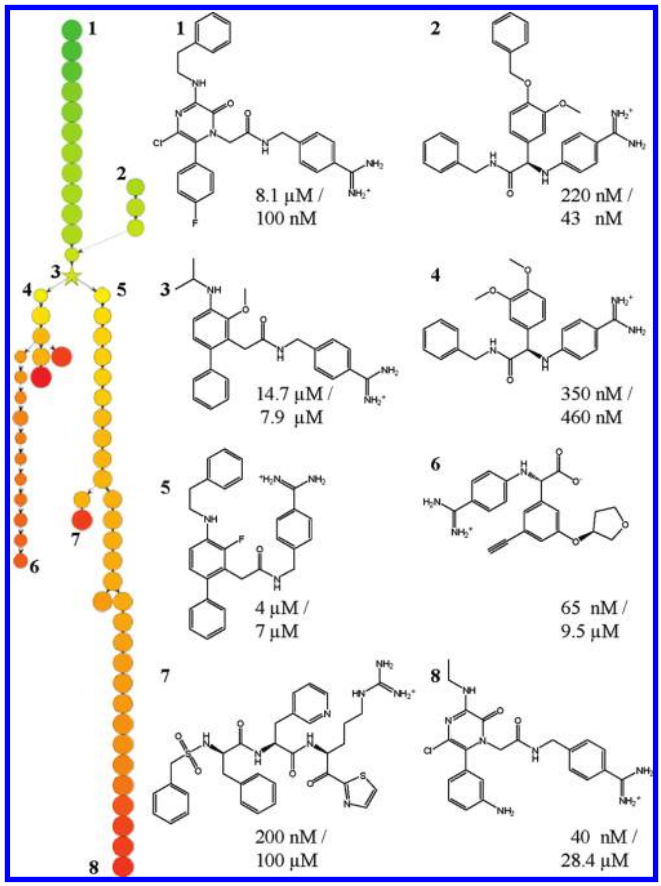


Figure 5. SSR tree. An SSR tree generated for a compound from the thrombin/factor VIIa selectivity set is shown. Green nodes represent compounds selective for thrombin and red nodes compounds selective for factor VIIa. The structures of eight compounds represented by individual nodes in the tree are shown (compound and node numbers correspond). The numbers below each compound report the potencies against thrombin/factor VIIa.

(see Results). Furthermore, different SAR and SSR pathways can be compared by synchronizing the selection of compounds.

Software Libraries. SARANEA is freely distributed under the GNU General Public License provided together with the program. It utilizes three public domain Java libraries: Java Universal Network/Graph Framework (JUNG2),[12] JChemPaint,[13] and JOELib.[14] Upon online publication of this paper, SARANEA and its source code, normalization and marker sets, compound data sets for exemplary applications, and program documentation can be obtained from http://www.lifescienceinformatics.uni-bonn.de ("Downloads" section).

Application to Serine Protease Inhibitors. SARANEA was applied to sets of inhibitors for three serine proteases (i.e., thrombin, plasmin, and factor VIIa) extracted from BindingDB[11] to analyze SARs and SSRs they form. In BindingDB and herein, compound potency is reported as K_i or IC_{50} values. For this application, a publicly available version of the MACCS fingerprint[15] was calculated for all compounds to assess molecular similarity. However, any binary fingerprint can be utilized in SARANEA (see the Supporting Information). The inhibitor and selectivity sets are summarized in Table 2, and the Supporting Information contains further details concerning the compound sets and fingerprint calculations.

RESULTS

Potency and Selectivity NSGs of Protease Inhibitors. SARANEA was applied to analyze SARs and SSRs for inhibitors of thrombin, plasmin, and factor VIIa. In addition to these protease inhibitor sets, other application examples have been presented in the original publications describing NSGs and SAR pathways. Furthermore, more than 300 exemplary sets of active compounds are provided together with the SARANEA package. From the three inhibitor sets, two target pair selectivity sets could be assembled (Table 2), and thus, three potency and two selectivity NSGs were generated. Thrombin inhibitors formed the largest set with 488 molecules that were characterized by a globally heterogeneous SAR. Thus, in this case, the coexistence of continuous and discontinuous SAR components is expected. This becomes apparent in Figure 1 when zooming in on the thrombin NSG, which reveals the presence of clusters with many small yellow nodes (i.e., structurally similar and also diverse compounds having similar potency, continuous SAR region) and clusters containing large red and green nodes that mark activity cliffs (i.e., similar inhibitors with large differences in potency, discontinuous SAR region). Different from thrombin, the plasmin inhibitor set produces a much higher continuity than discontinuity score and thus represents a globally continuous SAR phenotype. Figure 1 shows that most plasmin inhibitors indeed form a densely connected cluster with clear SAR continuity. Nevertheless, this inhibitor set also contains a smaller number of compounds that constitute a discontinuous cluster, and several of these compounds form notable activity cliffs. Thus, although the plasmin set is characterized by global SAR continuity, the NSG reveals that it also contains at least one activity cliff, hence illustrating the relationship between global and local SAR features. Furthermore, the two target pair selectivity sets yield lower SARI scores than the three inhibitor sets (Table 2). Accordingly, their selectivity NSGs in Figure 1 reveal the presence of multiple selectivity cliffs that are formed by structurally similar compounds with opposite target selectivity. Thus, potency and selectivity NSG representations provide a first view of global and local SAR and SSR features present in compound data sets that can be further refined using other SARANEA functions.

Characterization of Activity and Selectivity Cliffs. The next level of our analysis is a detailed characterization of activity and selectivity cliffs. Figure 2A shows three inhibitors with different potency and selectivity against thrombin and plasmin. Compounds 1 and 3, which form the dominant selectivity cliff in the thrombin/plasmin selectivity NSG, are selected together with a neighbor (compound 2). SARANEA automatically synchronizes this selection with the two corresponding potency NSGs, revealing that compounds 2 and 3 are equally potent against plasmin but have different potency against thrombin. The structures of these compounds show that small chemical modifications have opposite effects on the potency against these two targets, which provides a basis for target-selective optimization. Furthermore, we can also identify compound modifications that affect the potency against two targets in similar ways, as shown in Figure 2B for inhibitors of thrombin and factor VIIa. First, a cluster of analogues representing a substituted pyrazinone chemotype is selected from the thrombin/factor VIIa selectivity NSG

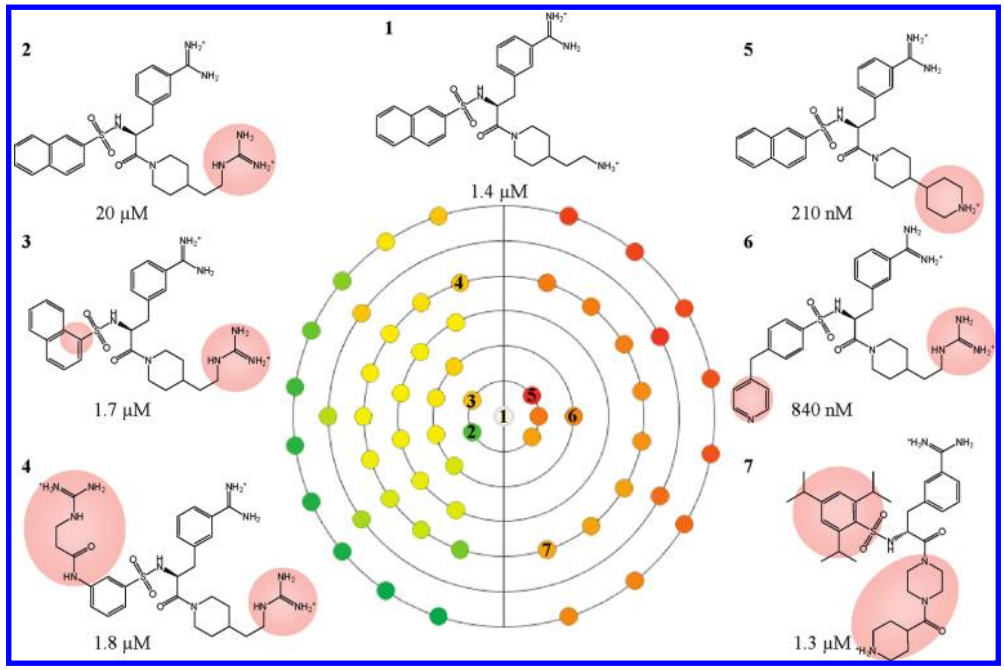


Figure 6. Exploring chemical neighborhoods. An exemplary layered chemical neighborhood graph is shown for plasmin inhibitors. The graph explores the neighborhood of compound 1. The neighbors of compound 1, which is placed in the center, are positioned on concentric circles that represent different levels of 2D structural similarity. Most similar compounds are found on the innermost circle, and nodes representing structurally increasingly dissimilar compounds are placed on subsequent peripheral circles. Nodes shown on the right side of the vertical line in the radial view represent compounds with higher potency than the central node, whereas compounds corresponding to nodes on the left side are less potent. Nodes are sorted and color-coded according to their potency, applying a continuous spectrum from green (least potent) to red (most potent). Compounds 2 and 5 represent a steep activity cliff because they are structurally very similar to the central molecule but show significantly lower (compound 2) or higher (compound 5) potency. Structural features of compounds discussed in the text are highlighted.

Table 2. Serine Protease Inhibitor and Target Pair Selectivity Sets[a]

no.	inhibitor set	cont	disc	SARI	no. of compds
1	thrombin	0.796	0.902	0.447	488
2	plasmin	0.626	0.200	0.713	96
3	factor VIIa	0.677	0.635	0.521	224
4	thrombin/plasmin	0.474	0.649	0.413	25
5	thrombin/factor VIIa	0.612	0.858	0.377	107

[a] Global SAR/SSR characteristics and compound numbers are reported for three inhibitor (1−3) and two selectivity (4, 5) sets. Plasmin and factor VIIa share only a single inhibitor, and thus, no selectivity set could be generated for this target pair. Abbreviations in column heads: cont, global continuity score; disc, global discontinuity score; no. of compds, number of compounds per set. The SARI values indicate the following global SAR/SSR categories[5,6] for these inhibitor sets: thrombin, factor VIIa, heterogeneous SAR; plasmin, continuous SAR; thrombin/plasmin, thrombin/factor VIIa, heterogeneous SSR.

(top left). Utilizing the table view of SARANEA, compounds having 2-fold or higher selectivity for one of the two targets are omitted. The nodes representing the remaining 14 compounds are interactively separated from the cluster through graph editing. For clarity and node accessibility, edges between nodes are not displayed in this case. Then the potency NSG for factor VIIa is generated, and the 14 compounds are also selected in this graph using the synchronization feature of SARANEA (and the corresponding nodes are also separated from the large graph component). As can be seen in Figure 2B, the selected compounds display a more differentiated potency than selectivity profile. In addition to color coding of nodes, the table view of SARANEA provides a quantitative assessment of potency and selectivity. From the potency NSG table, the compound with the highest potency for factor VIIa is selected (compound 1 in Figure 2B), and the remaining compounds in the isolated subset are added to the selection. This is done to sort the molecules on the basis of their cliff index in pairs with compound 1. Compound 2 produces the highest cliff index (5.65) and shows a nearly 100-fold lower potency than compound 1 for both factor VIIa and thrombin. To identify compounds with a minimal selectivity difference that form an activity cliff, we next select the second-most potent compound from the factor VIIa NSG table (compound 3 in Figure 2B) and synchronize the selection with the thrombin/factor VIIa selectivity NSG. From the selectivity NSG, neighbors of this compound are selected and sorted according to the selectivity cliff index, and the compound with the lowest cliff index (0.01) is chosen (compound 4 in Figure 2B). These two compounds show no difference in selectivity and are distinguished by only one structural modification, a substitution of a methyl group for a chlorine atom. However, compound 3 is 7 times more potent than compound 4 against both targets. These examples illustrate how activity and selectivity cliffs are analyzed with SARANEA and how compounds with a desired potency or selectivity profile are identified.

Compounds with Potential Side Effects. Sets of marker compounds having known undesired properties can be added to NSG calculations to identify active molecules that are structurally similar to markers and hence potentially problematic. For example, two carcinogenicity marker sets are added to the thrombin inhibitor potency NSG, applying a predefined similarity threshold value. The two marker sets consist of compounds known to induce cancer of the hematopoietic system and stomach, respectively. As shown

in Figure 3, a series of 6 thrombin inhibitors display distinct structural similarity to a total of 14 carcinogenic substances and are thus flagged for potential side effects.

Comparing and Editing SAR and SSR Pathways and Trees. To assess small structural changes that are accompanied by a gradual change in potency or selectivity, pathways are identified that consist of sequences of pairwise similar compounds forming an ascending potency gradient. Pathways organize SAR/SSR information available in screening data or other compound sets. For compound selection, pathways leading to activity cliffs are often of particular interest. Pathways originating from and/or leading to a selected compound are systematically computed (see the Supporting Information) and then interactively compared and edited. The SSR pathway shown in Figure 4 is selected using the SARTree function and edited by introducing "shortcuts" to reduce the number of compounds within the path. Then the compounds forming this pathway are identified in the corresponding selectivity NSG and the two underlying potency NSGs by synchronizing the selection. In the next step, the factor VIIa and thrombin SAR pathways containing the largest number of SSR pathway compounds are selected (parts B and C, respectively, of Figure 4). Comparison of these pathways reveals that the order of compounds in the SSR and SAR pathways is not identical because pairs of similar compounds having small selectivity differences can have large potency differences for each of the two targets.

The SSR pathway in Figure 4A covers compounds that are selective for factor VIIa (green nodes), have similar potency for factor VIIa and thrombin (yellow nodes), or are selective for thrombin (orange/red nodes). All compounds in the pathway share the benzenecarboximidamide group. However, compounds that are decreasingly selective for factor VIIa are found to contain a central substituted phenylglycine moiety that is replaced by a substituted benzene chloropyrazinone in compounds with increasing selectivity for thrombin. The factor VIIa SAR pathway in Figure 4B starts with weakly potent chloropyrazinone derivatives, which is consistent with their terminal positions in the SSR pathway, corresponding to high selectivity for thrombin over factor VIIa. The factor VIIa SAR pathway in Figure 4B and the thrombin SAR pathway in Figure 4C suggest that the phenylglycine moiety is important for achieving high potency against both targets. Thus, the phenylglycine group is a potency, but not selectivity, determinant in these inhibitors.

The SARTree utility of SARANEA also makes it possible to collect all pathways that contain a compound of interest and organize these pathways in an SAR/SSR tree data structure. The tree is rooted at the node representing the shared compound, and the leaves of the tree correspond to the start or end points of all pathways. Figure 5 shows an exemplary SSR tree for pathways sharing a compound from the thrombin/factorVIIa selectivity set. At the top, green nodes represent compounds selective for thrombin. Nodes 1 and 2 are start points of pathways that converge on the node representing the shared compound. At the bottom, red nodes (6, 7, and 8) represent pathway end points, i.e., compounds selective for factor VIIa over thrombin. Compounds 1 and 8 in Figure 5 constitute a selectivity cliff. They are neighbors in the selectivity NSG and have opposite selectivity for factor VIIa and thrombin. Because the SSR tree reflects gradual changes in selectivity, preferred chemotypes might be identified. For example, 39 of 42 compounds in the pathway between compounds 1 and 8 are chloropyrazinone analogues. By contrast, the pathway between compounds 2 and 6 mostly consists of phenylglycine derivatives. The SARTree view is complementary to the NSG view and permits the selection of different pathways from trees for editing and further analysis of the identified compound series.

Exploring the Chemical Neighborhood of Compounds. Layered chemical neighborhood graphs generated with SARANEA (as part of the pathway editing utility) explore the environment of selected compounds in detail. A radial view of the chemical neighborhood of a compound is produced that reveals activity or selectivity cliffs and also molecules that form continuous SARs or SSRs with the selected compound (i.e., molecules with decreasing structural similarity but comparable potency or selectivity). Figure 6 shows an example for plasmin inhibitors. Compounds are organized in layers around the central reference molecule. The nodes on each layer are sorted according to their potency. Thus, the color code and localization of nodes provide a basis for the selection of structurally similar compounds with large potency or selectivity differences or, alternatively, structurally dissimilar compounds with comparable potency or selectivity. In Figure 6, five compounds positioned on the innermost layer represent the immediate structural neighborhood of compound 1. However, despite their structural similarity, these compounds greatly differ in their potency against plasmin. The central compound 1 is only moderately potent (1.4 μM). In compound 2, the amino group is replaced by a guanidino group, which further reduces potency (20 μM). In compound 3, a change in the naphthalene attachment position essentially restores the potency (1.7 μM), despite the presence of the guanidino group. Three other compounds in the same similarity layer, positioned to the right of compound 1, are more potent, with compound 5 having the highest potency (210 nM), as indicated by its red color and its topmost position in the inner layer. This nearly 10-fold increase in potency relative to that of compound 1 is due to the introduction of a piperidine group. Compounds in different similarity layers are also compared. For example, compound 6 is positioned in the second and compounds 4 and 7 are positioned in the fourth layer of the graph. These compounds have potency values comparable to that of compound 1. However, the structural deviations are larger in these cases. Compound 6 has a 2-benzylpiperidine moiety attached to the sulfoxy group instead of the naphthalene. Compound 4 contains two guanidino groups, and compound 7 substantially differs from compound 1 but has comparable potency (1.3 μM). These examples illustrate how chemical neighborhood graphs are used to analyze structural determinants of compound potency or selectivity.

Comparison to Other Software Tools. We have compared SARANEA's major features to those of four other software packages, i.e., DrugViz,[16] ChemGPS,[17] Scaffold Hunter,[18] and enhanced SAR maps.[19]

DrugViz is a plug-in for Cytoscape,[20] a Java program primarily used for visualization of biological interaction networks. DrugViz adds ligands as nodes to target interaction networks and allows the connection of targets based on common ligands. It is also possible to find molecules that are similar to a particular ligand using fingerprint overlap

or graph similarity metrics. However, molecular similarity is not applied to build the network, and potency information is not directly incorporated into the data structure. Rather, shared ligands and target interactions define the network. The identification of activity or selectivity cliff markers is not possible.

ChemGPS is available as a Web service and projects molecules into a low-dimensional chemical descriptor space. The chemical space representation has been optimized to reflect chemical properties relevant for bioactivity. Molecules can then be compared and features identified that are associated with activity. However, activity annotation is not used to generate the space representation. By contrast, SARANEA relates molecular similarity directly to activity and selectivity.

Scaffold Hunter uses a hierarchical fragmentation scheme, the so-called scaffold tree, to identify different scaffolds that represent active compounds. The tree reflects the hierarchical fragmentation procedure and allows the identification of scaffolds that have not been covered by molecules in a given data set. Individual scaffolds can be annotated with potency information of ligands they represent. Therefore, individual compounds in Scaffold Hunter are compared on the basis of common scaffolds, rather than overall molecular similarity. While scaffolds that are associated with high bioactivity can be identified, the tree structure does not incorporate SAR characteristics such as continuity and discontinuity.

Enhanced SAR maps show the distribution of ligand properties with respect to different functional groups. This allows the identification of structural features that are associated with certain biological activities or other properties. In contrast to SARANEA, molecules are compared on the basis of R-group decomposition. While this approach is well-suited for analogue series, it is not directly applicable to heterogeneous compound sets. By contrast, in SARANEA, NSGs capture an entire compound set. Moreover, SAR or SSR pathways can be identified that lead from one analogue series to another. However, the identification of structural features that are characteristic of activity cliff marker compounds is not a part of the SARANEA program.

In summary, SARANEA is complementary to other SAR analysis programs and approaches. It is unique in that it utilizes molecular similarity and activity data to build a network representation of the data set. This makes it possible to systematically explore and identify activity or selectivity cliffs and their markers. These key compounds and/or molecules forming SAR or SSR pathways can then be further analyzed.

DISCUSSION

SARANEA allows the systematic analysis of SARs and SSRs across multiple activity classes. The central data structure utilized in the program is the NSG, which distinguishes SARANEA from other SAR analysis or drug−target network programs. Different from drug−target networks, where small molecules are compared on the basis of common targets, SARANEA builds molecular similarity-based network representations for a given data set. Similarity relationships between molecules are directly visualized in NSGs and allow the identification of SAR pathways. Therefore, the program is complementary to other available SAR analysis and network generating tools.

SARANEA is made freely available with its full source code to enable the modification, extension, and further development of its tools in the scientific community. As provided, the program can be applied in combination with different fingerprints that need to be calculated externally. It is conceivable that fingerprinting modules might further extend future versions of SARANEA, so that different molecular representations can be directly compared. Furthermore, 2D depictions of molecules are currently calculated automatically from SMILES strings. Hence, another possible extension might be to accept atom coordinate formats such that the user can display molecules with more sophisticated methods.

From a methodological point of view, SARANEA integrates and further extends SAR analysis approaches recently developed by us. As we have shown, the applications of fingerprints and Tanimoto similarity produce meaningful NSGs. However, it might also be of interest to evaluate other descriptor-based molecular representations and similarity metrics that are currently not implemented.

To permit the analysis of multiple SARs and SSRs in parallel, we have chosen a selection synchronization approach, as described above, which represents a straightforward way to evaluate SARs and SSRs of different targets. However, one might also incorporate more than two potency values into one NSG to profile multiple targets. This feature is currently not available but might be implemented, for example, by using pie charts as nodes. Targets in SARANEA are currently available as a drop-down selection list that also reports the number of ligands shared by any two targets. This information could in principle also be visualized in a target network. In such a network representation, each node would correspond to a target and the associated SAR data set, while the edges would represent SSR data associated with a given target pair.

SARANEA can be easily extended by adding new modules. Furthermore, the calculation of NSGs is implemented as a visualization-independent module and might thus also be incorporated into other software packages.

CONCLUSIONS

The analysis of structure−activity relationships is generally challenged by increasing amounts of bioactivity data that need to be considered. Furthermore, how to consistently extract SAR information from compound and screening data sets still represents a largely unsolved problem in medicinal chemistry. Accordingly, there is a growing need for data-driven approaches that make it possible to study SARs on a large scale and in a comparative manner. Also, evidence is accumulating that many compounds originally thought to be target-specific do act on multiple targets, albeit with different selectivity. Thus, the study of molecular selectivity naturally complements SAR analysis but requires a methodological framework to dissect multitarget SARs for series of active compounds. SARANEA is designed to integrate different SAR and SSR analysis functions and enable systematic mining of structure−activity and structure−selectivity relationships in compound data sets of any source. The analysis features are data-oriented and intuitive visualization tools that

provide complementary views of molecular similarity, and property distributions play a central role for interactive analysis. The activity of compounds against multiple targets can be studied at different levels, and activity or selectivity cliffs can be easily identified and analyzed in detail. The application of SARANEA to serine protease inhibitor sets presented herein revealed the presence of in part overlapping yet distinct structural determinants of compound potency and selectivity against multiple targets. SARANEA is freely available as an easy to use and documented Java implementation including its source code. We hope that the program will be widely used and further extended to advance the SAR analysis field.

Supporting Information Available: Information about the theory underlying all analysis functions implemented in SARANEA, calculation details, and compound data set information. This information is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Rognan, D. Chemogenomic Approaches to Rational Drug Design. *Br. J. Pharmacol.* **2007**, *152*, 38–52.
(2) Peltason, L.; Bajorath, J. Systematic Computational Analysis of Structure−Activity Relationships: Concepts, Challenges, and Recent Advances. *Future Med. Chem.* **2009**, *1*, 451–466.
(3) 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.
(4) Bajorath, J. Computational Analysis of Ligand Relationships within Target Families. *Curr. Opin. Chem. Biol.* **2008**, *12*, 352–358.
(5) Peltason, L.; Hu, Y.; Bajorath, J. From Structure−Activity to Structure−Selectivity Relationships: Quantitative Assessment, Selectivity Cliffs, and Key Compounds. *ChemMedChem* **2009**, *4*, 1864–1873.
(6) Peltason, L.; Bajorath, J. SAR Index: Quantifying the Nature of Structure−Activity Relationships. *J. Med. Chem.* **2007**, *50*, 5571–5578.
(7) 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.
(8) Wawer, M.; Peltason, L.; Bajorath, J. Elucidation of Structure−Activity Relationship Pathways in Biological Screening Data. *J. Med. Chem.* **2009**, *52*, 1075–1080.
(9) Wawer, M.; Bajorath, J. Systematic Extraction of Structure−Activity Relationship Information from Biological Screening Data. *ChemMedChem* **2009**, *4*, 1431–1438.
(10) Gold, L. S.; Sawyer, C. B.; Magaw, R.; Backman, G. M.; de Veciana, M.; Levinson, R.; Hooper, N. K.; Havender, W. R.; Bernstein, L.; Peto, R.; Pike, M. C.; Ames, B. N. A Carcinogenic Potency Database of the Standardized Results of Animal Bioassays. *Environ. Health Perspect.* **1984**, *58*, 9–319.
(11) 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.
(12) Java Universal Network/Graph Framework. http://jung.sourceforge.net/ (accessed June 1, 2009).
(13) Krause, S.; Willighagen, E.; Steinbeck, C. JChemPaint−Using the Collaborative Forces of the Internet To Develop a Free Editor for 2D Chemical Structures. *Molecules* **2000**, *5*, 93–98.
(14) JOELib: A Java Based Computational chemistry Package. http://joelib.sourceforge.net/ (accessed Aug 10, 2009).
(15) *MACCS Structural Keys*; Symyx Software: San Ramon, CA, 2005.
(16) Xiong, B.; Liu, K.; Wu, J.; Burk, D. L.; Jiang, H.; Shen, J. DrugViz: A Cytoscape Plugin for Visualizing and Analyzing Small Molecule Drugs in Biological Networks. *Bioinformatics* **2008**, *24*, 2117–2118.
(17) Larsson, J.; Gottfries, J.; Bohlin, L.; Backlund, A. Expanding the ChemGPS Chemical Space with Natural Products. *J. Nat. Prod.* **2005**, *68*, 985–991.
(18) Wetzel, S.; Klein, K.; Renner, S.; Rauh, D.; Oprea, T. I.; Mutzel, P.; Waldmann, H. Interactive Exploration of Chemical Space with Scaffold Hunter. *Nat. Chem. Biol.* **2009**, *5*, 581–583.
(19) 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.
(20) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **2003**, *13*, 2498–2504.

CI900416A